PAC Guarantees for Reinforcement Learning: Sample Complexity, Coverage, and Structure

Joshua Steier joshsteier@gmail.com
Independent Researcher

October 21, 2025

Abstract

Fixed–confidence (PAC) guarantees are the right primitive when data are scarce or failures are costly. This survey organizes the 2018–2025 literature through a Coverage–Structure–Objective (CSO) template, in which sample complexity satisfies $N(\varepsilon, \delta) \approx \mathsf{Cov} \times \mathsf{Comp} \times \mathrm{poly}(H) \times \varepsilon^{-2}$. Coverage captures access assumptions (online/generative vs. offline via concentrability); Structure captures problem–dependent capacity (tabular SA, linear dimension d, effective dimension $d_{\mathrm{eff}}(\lambda)$, rank r, Bellman/witness/BE measures); Objective fixes the target (uniform–PAC/regret, instance–dependent identification, reward–free exploration, offline control/OPE, partial observability). We synthesize: tight tabular baselines; the uniform–PAC $\Rightarrow$ high–probability regret bridge; structured learnability under Bellman rank and Bellman–Eluder dimension; linear, kernel/NTK, and low–rank models; reward–free exploration as coverage creation; and pessimistic offline RL with explicit coverage dependence. Practical outputs include a rate "cookbook," a decision tree, and a unified roadmap of open problems (kernel/NTK uniform–PAC, agnostic low–rank, misspecified offline RL, instance–dependent FA, structure selection). We unify notation, state results with explicit dependencies, and provide a decision toolkit for practitioners.

Keywords PAC RL; uniform–PAC; sample complexity; function approximation; offline RL; low–rank MDPs; Bellman–Eluder dimension.

How to read this survey. Readers mainly interested in *what to use when* can read Table 2, the CSO template (§2), and the practical decision tree (§15); then jump to the relevant setting: tabular baselines (§4), structural complexity measures (§5), function approximation (§6), rich observations/low rank (§8), reward-free exploration (§9), offline RL (§10), partial observability (§11), or PAC–Bayes (§12). Readers seeking *technical underpinnings* should read §3 and the definition blocks at the start of each section. We keep proofs informal (dependencies and scalings) and point to the primary sources for exact constants.

Figures 1 and 2, together with Tables 2 and 5, provide a quick triage guide.

1 Introduction

Reinforcement learning in brief. In reinforcement learning (RL), an agent interacts with an environment over discrete time steps. At each step, the agent observes a *state* s (e.g., a robot's sensor readings), selects an *action* a (e.g., a motor command), receives a scalar *reward* r (e.g., $+1$ for reaching a goal), and transitions to a new state s'. The agent's objective is to learn a *policy*—a rule π mapping states to actions—that maximizes cumulative reward over a planning horizon H. The central challenge is the *exploration–exploitation tradeoff*: the agent must try unfamiliar actions to discover their consequences while also exploiting actions already known to be good.

This survey focuses on *fixed-confidence* (PAC) guarantees: formal bounds ensuring that, with probability at least $1 - \delta$, the learned policy is ε-optimal after a specified number of interactions. Such guarantees are essential when data collection is expensive or when failures are costly—settings where average-case performance is insufficient and worst-case assurances are required.

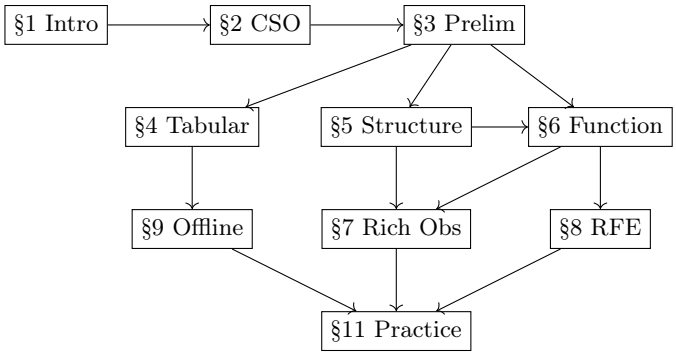

Figure 1: **Survey roadmap.** CSO framework (§2) organizes all results. Preliminaries (§3) introduce formal tools. Three pillars—tabular (§4), structural measures (§5), function approximation (§6)—feed into applications (§7–§9). Practice (§11) synthesizes. **Reading paths:** Practitioners → §2, §11, then domain sections; theorists → §3, then §4–§6.

1. Motivation & precise scope. This survey asks: *what fixed–confidence (PAC) guarantees are known for reinforcement learning (RL), which assumptions make them possible, and how do the guarantees scale with problem parameters?* Fixed–confidence targets are essential in safety–critical and data–scarce regimes because they bound the probability of unacceptable performance, in contrast to average regret. From 2018–2025, several strands matured: (i) *uniform–PAC* bounds that connect PAC guarantees to high–probability regret (Dann et al., 2017; [2]); (ii) *instance–dependent* identification beyond worst–case rates (Wagenmaker et al., 2022; Tirinzoni et al., 2023; [33, 32]); (iii) *coverage–centric* formulations such as reward–free exploration (Jin et al., 2020; [15]) and pessimistic offline RL (Jin–Yang–Wang, 2021; Shi et al., 2022; [18, 26]); and (iv) *function–approximation* regimes characterized by structural measures (Jiang et al., 2017; Jin–Liu–Miryoosefi, 2021; [14, 17]). This article synthesizes those results, emphasizing sample complexity (PAC and uniform–PAC), exploration, function approximation (linear, low–rank, kernel/over–parameterized), offline RL, partial observability, and structural risk measures. Scope is restricted to 2018–2025, with pre–2018 foundations directed to Strehl–Li–Littman (2009; [27]), Szepesvári (2010; [30]), Kaelbling–Littman–Moore (1996; [19]), and Sutton–Barto (2018; [29]).

Scope window (2018–2025). We focus on 2018–2025 because: (i) uniform–PAC crystallized the PAC↔regret bridge in late 2017 [2]; (ii) structural complexity measures (Bellman rank, witness rank, Bellman–Eluder) matured 2017–2021 [14, 17]; and (iii) coverage–centric offline RL (pessimism, concentrability) emerged post–2020 with sharp characterizations by 2024 [18, 26, 22, 36].

2. Core concepts (informal; formalized in §3). Let $M = (\mathcal{S}, \mathcal{A}, P, r, H, \rho)$ be a finite–horizon MDP with $|\mathcal{S}| = S$, $|\mathcal{A}| = A$, rewards in $[0, 1]$, horizon $H \in \mathbb{N}$, and start–state distribution ρ. For a policy π, write $V_1^\pi(\rho) = \mathbb{E}_{s\sim\rho}[V_1^\pi(s)]$.

An algorithm is (ε, δ)–**PAC** if, with probability $\geq 1 - \delta$, it returns an ε-optimal policy $\hat{\pi}$ after $N(\varepsilon, \delta)$ episodes (Definition 1). **Uniform–PAC** bounds the number of ε-suboptimal episodes for *all* ε simultaneously, implying high-probability regret (Definition 2, Theorem 5). Key structural parameters—Bellman rank B, witness rank W, Bellman–Eluder dimension d_{BE}, feature dimension d, rank r—replace (S, A) in sample complexity bounds (§5, §6). **Coverage** (quantified by concentrability $C_\star$) determines offline RL feasibility (Definition 6, §10).

3. Canonical results.

Uniform–PAC ⇒ regret bridge (Theorem 5, informal). If an algorithm is uniform–PAC with budget $N(\varepsilon, \delta)$, then its cumulative regret satisfies $\mathrm{Regret}(K) = \mathcal{O}\big(\int_0^H \min\{K,\, N(\varepsilon, \delta)\}\, d\varepsilon\big)$ with high probability. This bridge, formalized in §3, recovers near-minimax tabular regret when $N(\varepsilon, \delta)$ has polynomial dependence in $(S, A, H, 1/\varepsilon)$ [2].

Theorem 1 (Reward–free exploration (tabular))**.** *There exist algorithms that, without access to rewards, collect* $\tilde{\mathcal{O}}\big(S^2 A\, H^3\, \varepsilon^{-2}\big)$ *episodes and subsequently, for any reward function in* $[0,1]$*, output an* ε*–optimal policy with probability* $1-\delta$ *[15].*

Theorem 2 (Linear MDPs)**.** *Under linear transition/reward realizability with feature dimension* d*, there exist polynomial-time algorithms with* $N(\varepsilon,\delta) = \tilde{\mathcal{O}}\big(d^3 H^5 \varepsilon^{-2}\big)$ *(see uniform–PAC refinements [11]) and* $\mathrm{Regret}(K) = \tilde{\mathcal{O}}\big(d\sqrt{H^3 K}\big)$*. See Theorem 8 for details and constants [16].*

Theorem 3 (Function classes with bounded Bellman–Eluder dimension)**.** *If the Bellman–Eluder dimension* d_{BE} *of the value class is finite, then sample complexity/regret bounds scale polynomially in* (d_{BE}, H) *with logarithmic dependence on confidence (Jin–Liu–Miryoosefi, 2021; [17]).*

Theorem 4 (Offline RL via pessimism)**.** *In offline RL with linear MDP realizability and concentrability* $C_\star$*, pessimistic algorithms such as PQL attain value–estimation and control error* ε *with data size* $\tilde{\mathcal{O}}\big(\mathrm{poly}(H)\, d\, \mathrm{poly}(C_\star)\, \varepsilon^{-2}\big)$ *(Shi et al., 2022; [26]); model–based counterparts can be minimax–optimal in tabular settings (Li et al., 2024; [22]).*

Notes. Precise constants and horizon exponents appear in the cited theorems; we state dependencies to orient the reader. Where exponents are sensitive to modeling choices, we flag refinements later.

4. Connections to other results.

- **Uniform–PAC as a bridge**: Uniform–PAC converts fixed–confidence guarantees into high–probability regret, aligning PAC targets with learning–curve behavior [2].
- **Coverage as a unifier**: Reward–free exploration and offline pessimism both hinge on coverage—either constructed online or inherited from a behavior policy [15, 26].
- **Structure ⇒ tractability**: Bellman rank and Bellman–Eluder dimension delineate when function approximation admits PAC–style learning [14, 17].
- **Instance–dependence**: Identification can be strictly easier than worst–case learning; gap–based analyses refine $\tilde{\mathcal{O}}$ rates [33, 32].

5. Practical implications.

- Verify realizability or adopt algorithms robust to misspecification; avoid applying linear–MDP guarantees to arbitrary deep networks without structure.
- For *multi–task* settings, consider reward–free exploration to amortize exploration across rewards [15].
- In *offline* pipelines, quantify coverage (e.g., surrogates for $C_\star$) and prefer pessimism; avoid deployment when estimated coverage is poor [26].
- Use policy certificates where available to obtain per–episode accountability (Dann et al., 2019; [4]).

6. Open problems.

i) Uniform–PAC with kernel/over–parameterized classes under minimal spectrum assumptions; target $\tilde{\mathcal{O}}(d_{\mathrm{eff}}, H, \varepsilon^{-2})$ rates with computational efficiency.

ii) Agnostic low–rank/latent–state PAC guarantees with polynomial dependence in rank/latent size and horizon.

iii) Offline RL under model misspecification: bounds that separate estimation error, coverage, and approximation error with sharp trade–offs.

iv) Data–dependent structural selection (SRM in RL): oracle inequalities that adapt to Bellman–Eluder/Bellman rank estimated from data.

	Szepesvári (2010)	Strehl et al. (2009)	This survey (2018–2025)
Venue/Type	Monograph (Morgan & Claypool, Synthesis Lectures)	JMLR survey article	TMLR survey (2025)
Scope	Algorithms for RL; primarily tabular and basic function approximation	PAC-MDP framework; finite MDPs; sample complexity	Uniform PAC/PAC across tabular, structural measures, FA, RFE, offline, POMDP
Time window	Pre-2010	Pre-2009	2018–2025 (uniform-PAC era)
Primary axis	Algorithms & theory (value iteration, TD, policy gradient)	PAC-MDP sample complexity	CSO: Coverage × Structure × Objective
Guarantee types	PAC/Regret (primarily tabular); some FA	PAC-MDP bounds for finite state-action spaces	Uniform-PAC ↔ regret; instance-dependent; offline/pessimism; PAC–Bayes
Practitioner tools	Conceptual algorithms	Theoretical foundations	Decision tree; coverage gates; certificates; rate cookbook
What's new here	N/A	N/A	Uniform-PAC synthesis; coverage-centric view; structural measures side-by-side; deployment gates

Table 1: **Comparison with foundational surveys.** We build on Szepesvári's algorithmic treatment and Strehl et al.'s PAC-MDP framework, adding 2018–2025 advances under a unified CSO lens with deployment-oriented tools.

7. Mini–bibliography (context in one–two sentences each). Strehl–Li–Littman (2009; [27]) surveyed PAC–MDP analyses in finite MDPs and set classical baselines for $(S, A, H, \varepsilon, \delta)$–dependent sample complexity. Szepesvári (2010; [30]) provided a monograph–level treatment of RL algorithms and analysis. Kaelbling–Littman–Moore (1996; [19]) offered an early broad survey of RL methods; Sutton–Barto (2018; [29]) updated foundational algorithmics. Dann–Lattimore–Brunskill (2017; [2]) introduced uniform–PAC, linking fixed–confidence and regret in episodic RL. Jiang et al. (2017; [14]) formalized contextual decision processes and Bellman rank; Jin–Liu–Miryoosefi (2021; [17]) introduced Bellman–Eluder dimension. Jin et al. (2020; [16]) established provable learning with linear function approximation; He–Zhou–Gu (2021; [11]) derived uniform–PAC in linear MDPs. Jin et al. (2020; [15]) formalized reward–free exploration with near–optimal dependence on $(S, A, H, 1/\varepsilon)$. Dann et al. (2019; [4]) proposed policy certificates for accountable RL. Jin–Yang–Wang (2021; [18]) and Shi et al. (2022; [26]) developed pessimistic offline RL with coverage–dependent guarantees; Li et al. (2024; [22]) settled model–based offline sample complexity in tabular regimes. Wagenmaker et al. (2022; [33]) and Tirinzoni et al. (2023; [32]) advanced instance–dependent PAC identification beyond worst–case rates.

Related surveys and how we differ. Szepesvári's monograph [30] and Strehl–Li–Littman [27] cover classical PAC–MDP and tabular RL. Our contribution is a *unified, book-length synthesis* across 2018–2025 that (i) connects PAC and regret via uniform–PAC across settings, (ii) places *coverage* at center stage (RFE vs. offline), (iii) systematizes structural measures (Bellman/witness/BE/bilinear) side-by-side, and (iv) provides practitioner-grade toolkits (coverage estimation, certificates) with deployment gates.

2 Unified Assumption–Guarantee Framework & Taxonomy

Axes. We organise PAC–style results for RL along four axes and summarise each theorem as a 4–tuple $(\text{access}, \text{structure}, \text{objective}, \text{rate})$:

- **Access & coverage:** online interaction, generative model, or offline dataset. Coverage is quantified by coefficients (e.g., $C_\star$) that bound how well the data support the target policy.
- **Structure (complexity):** tabular; linear features (d); kernel/RKHS or NTK via effective dimension $d_{\text{eff}}(\lambda)$; low–rank/latent; or measures such as Bellman rank (B), witness rank (W), Bellman–Eluder

Access & coverage	Structure (param)	Objective	Representative guarantee & refs.	Computation / oracles
Online	Tabular	Uniform–PAC $\leftrightarrow$ Regret	$N = \tilde{\Theta}(SAH^3/\varepsilon^2)$; regret $\tilde{\Theta}(\sqrt{SAH^3K})$ [6, 36, 2].	Value iteration; poly-time.
Online	Tabular	Instance-dependent BPI	$\tilde{O}(\sum_{h,s} q_h(s) \sum_{a:\Delta_h(s,a)>\varepsilon} \Delta_h(s,a)^{-2})$ [33, 32].	UCB-style; poly-time.
Online (no rewards)	Tabular	Reward-free	$\tilde{O}(S^2AH^3/\varepsilon^2)$ [15].	Model-based or CI; poly-time.
Online	Linear MDP (d)	Uniform–PAC / regret	$N = \tilde{O}(d^3H^5/\varepsilon^2)$ (to H^4 with refinements); regret $\tilde{O}(d\sqrt{H^3K})$ [16, 11].	LSVI + ridge; poly-time.
Online / gen. model	Kernel/NTK (d_{eff})	PAC / regret	$\tilde{O}(\text{poly}(H)\, d_{\text{eff}}/\varepsilon^2)$; regret $\tilde{O}(\text{poly}(H)\sqrt{d_{\text{eff}}K})$ under completeness [35].	Kernel ridge; poly-time.
Online / generative	Bellman rank B, Witness rank W, BE d_{BE}, Bilinear r	PAC / regret	$\tilde{O}(\text{poly}(H,B)\varepsilon^{-2})$, $\tilde{O}(\text{poly}(H,W)\varepsilon^{-2})$, or polynomial in d_{BE}, r [14, 28, 17, 8].	Often oracle-based (CLS/CSL).
Online / generative	Low-rank / latent (r or m)	PAC / exploration	$\tilde{O}(\text{poly}(H)\text{poly}(r)\varepsilon^{-2})$ (or in m) under identifiability/decodability [7, 1, 5, 12].	Rep. learning + planning.
Offline ($C_\star$)	Tabular / Linear	Pessimistic control; OPE	$\tilde{O}(\text{poly}(H)\,\text{poly}(\text{Comp})\,\text{poly}(C_\star)\,\varepsilon^{-2})$; tabular model-based minimax [22]; efficient OPE [20, 13].	PEVI/PQL; OPE DR/AIPW; poly-time.
Any	Policy class (PAC–Bayes)	Eval / selection	KL-penalized bounds for randomized policies with unbounded losses [9, 24, 10, 31].	Convex opt; poly-time.

Table 2: Taxonomy with computational viewpoint. $\tilde{O}$ suppresses polylog factors. "CLS/CSL" = cost-sensitive classification/regression oracles.

dimension (d_{BE}), bilinear rank (r).

- **Objective:** uniform–PAC/regret; instance–dependent identification (BPI); reward–free exploration; offline control or OPE; partial observability; PAC–Bayes evaluation/selection.
- **Rate:** the leading sample complexity/regret dependence in (H, ε, δ) and the structural parameter(s).

CSO template. Most PAC results in this survey can be read through a single template:

$$N(\varepsilon,\delta) \;=\; \tilde{\mathcal{O}}\Big(\underbrace{\text{poly}(H)}_{\text{planning depth}} \cdot \underbrace{\mathsf{Comp}(\text{structure})}_{\text{capacity}} \cdot \underbrace{\mathsf{Cov}(\text{access/coverage})}_{\text{support}} \cdot \varepsilon^{-2} \cdot \log(1/\delta)\Big)$$

Assumptions and Scope. The guarantees in this survey hold *only when structural assumptions are satisfied*:

- **Realizability**: $Q^\star \in \mathcal{F}$ or $V^\star \in \mathcal{F}$—often violated by neural networks without careful architecture design.
- **Bellman completeness**: $\mathcal{T}_h\mathcal{F} \subseteq \mathcal{F}$—rarely verified in practice; violations cause unbounded error propagation.
- **Coverage**: Offline guarantees require bounded concentrability $C_\star < \infty$; violations cause silent failures where policies appear good on data but fail on deployment.

Before deployment: Use the diagnostics in §15 (Algorithms 1–2) to verify assumptions. When assumptions are uncertain, prefer pessimism and interval-valued evaluation over point policy improvement.

Setting	Sample Complexity	H Exponent	Tightness
Tabular (online)	$\tilde{\Theta}(SAH^3/\varepsilon^2)$	H^3	Tight [36]
Linear MDP	$\tilde{O}(d^3H^5/\varepsilon^2)$ to $\tilde{O}(d^2H^4/\varepsilon^2)$	H^4–H^5	Upper; H^4 with variance reduction
Kernel/NTK	$\tilde{O}(d_{\text{eff}}(\lambda) \cdot H^4/\varepsilon^2)$ to $\tilde{O}(d_{\text{eff}}(\lambda) \cdot H^6/\varepsilon^2)$	H^4–H^6	Upper; lower bounds open
Low-rank (r)	$\tilde{O}(\text{poly}(r) \cdot H^3/\varepsilon^2)$ to $\tilde{O}(\text{poly}(r) \cdot H^4/\varepsilon^2)$	H^3–H^4	Upper; depends on identifiability
Reward-free (tabular)	$\tilde{O}(S^2AH^3/\varepsilon^2)$	H^3	Tight [15]
Reward-free (linear)	$\tilde{O}(d^2H^5/\varepsilon^2)$	H^5	Upper [34]
Offline (tabular)	$\tilde{\Theta}(SC_\star H^3/\varepsilon^2)$	H^3	Tight [22]
Offline (linear)	$\tilde{O}(d \cdot C_\star \cdot H^4/\varepsilon^2)$	H^4	Upper [26]

Table 3: **Horizon exponents across settings.** "Tight" indicates matching upper and lower bounds (up to logs). "Upper" indicates that only upper bounds are known; optimal H exponent remains open. The H^3 tabular exponent arises from variance accumulation over the horizon; higher exponents in function approximation settings reflect error compounding through correlated Bellman backups.

How settings relate. The structural settings in this survey form a hierarchy: results in more general settings recover special cases when parameters degenerate. Understanding these connections clarifies when guarantees transfer and why rates differ.

Tabular as a special case. Every structured setting recovers tabular MDPs under appropriate parameter choices:

- **Linear MDP** with one-hot features ($d = SA$) recovers tabular rates $\tilde{O}(SAH^3/\varepsilon^2)$.
- **Bellman rank** $B = S$ with indicator features over states.
- **Witness rank** $W = S$ when the model class contains all tabular MDPs.
- **Low-rank** with $r = S$ recovers tabular (the transition matrix has rank at most S).
- **Bellman–Eluder dimension** $d_{\text{BE}} = SA$ for tabular value classes.

Inclusion hierarchy. The settings satisfy the following (strict) inclusions, where each arrow denotes "is a special case of":

$$\text{Tabular} \subset \text{Linear MDP} \subset \text{Low-rank MDP} \subset \text{Bilinear class} \subset \text{Finite } d_{\text{BE}}$$

Each inclusion is strict: there exist problems in the larger class with infinite complexity in the smaller class. For example, continuous state spaces have $S = \infty$ but can have finite d (linear) or finite r (low-rank). Block MDPs with latent states $m \ll S$ are a special case of low-rank MDPs with $r = m$.

When do bounds match across settings?

- Linear MDP bounds *recover* tabular bounds when $d = SA$, but with potentially worse H exponents (H^4–H^5 vs. H^3) due to correlated estimation errors.
- Bellman rank and witness rank bounds are *incomparable*: neither subsumes the other in general, but both recover tabular with $B = W = S$.
- Bellman–Eluder dimension provides the most general guarantee but with less explicit dependence on problem parameters; it subsumes all other settings when $d_{\text{BE}} < \infty$.

Practical guidance. When analyzing a new problem:

1. First check if the problem is tabular (S, A finite and small); if so, use tabular bounds.
2. If S is large but features exist, check linear realizability; use linear MDP bounds with $d \ll SA$.

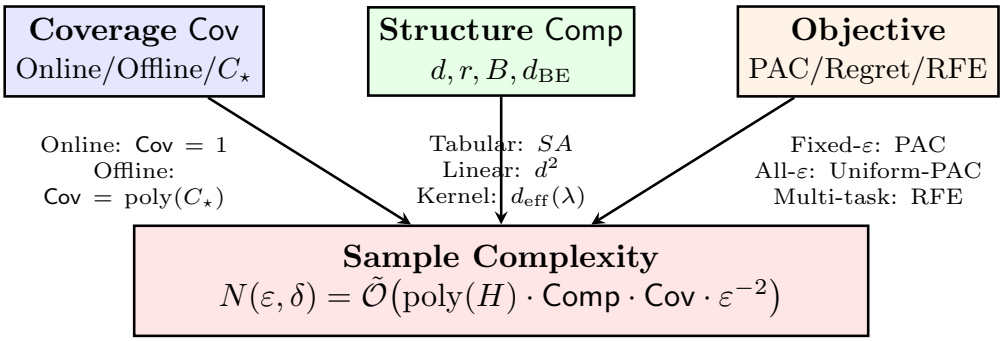

Figure 2: **CSO template: three knobs for sample complexity.** Every PAC result decomposes into Coverage × Structure × Objective. **Key insight:** the knobs are independent—improve any one to reduce $N(\varepsilon, \delta)$. **Practice:** if N is large, diagnose which knob is binding: (i) coverage too weak (collect more/go online), (ii) structure too complex (better features/representation), or (iii) mismatched objective (use reward-free for multi-task). **Theory:** lower bounds hold when the other two knobs are fixed; improving one does not contradict another's lower bound.

3. If transitions factor through latent states, check low-rank or Block MDP structure.
4. For general function classes, compute or bound d_{BE} to obtain guarantees.

Why horizon exponents differ. In tabular settings, concentration inequalities apply independently per state-action pair, yielding H^3 from the product of horizon length (H), variance accumulation (H), and the union bound over stages (H). In linear and kernel settings, estimation errors are correlated across the feature space: a single poorly-estimated direction affects all states sharing that feature, causing errors to compound multiplicatively through Bellman backups. This yields H^4–H^5 for linear MDPs, with the gap depending on whether variance-aware estimators are used. Kernel/NTK settings inherit similar compounding, with additional dependence on spectral properties of the kernel operator. Tighter exponents require either variance-reduced estimators, instance-dependent analyses, or fundamentally different algorithmic approaches.

- **Coverage Cov (access):** $\mathsf{Cov} = 1$ online (sufficient exploration) or with a generative model; $\mathsf{Cov} = \mathrm{poly}(C_\star)$ offline; and in reward–free exploration, Cov is the exploration budget that *creates* reusable support.
- **Structure Comp (capacity):** tabular $\Rightarrow \Phi_{\mathrm{tab}}(S, A, H)$; linear MDPs $\Rightarrow \mathrm{poly}(d)$; kernel/NTK $\Rightarrow d_{\mathrm{eff}}(\lambda)$; Bellman/witness ranks $\Rightarrow \mathrm{poly}(B)$ or $\mathrm{poly}(W)$; BE dimension $\Rightarrow \mathrm{poly}(d_{\mathrm{BE}})$; bilinear classes $\Rightarrow \mathrm{poly}(r)$.
- **Objective:** uniform–PAC/regret, instance–dependent identification, reward–free, offline control/OPE, POMDP/latent.

This *CSO* view makes explicit which knob the practitioner must turn in each setting: obtain/estimate coverage, certify structure (or use structural measures), then choose the objective and read off $N(\varepsilon, \delta)$ from the corresponding Comp and Cov.

Decision recipe (operational). Step 1 (access/coverage): Online or generative ⇒ proceed; Offline ⇒ estimate coverage proxies for $C_\star$ and prefer pessimism when coverage is limited. **Step 2 (structure):** Verify tabular/linear/kernel/low-rank assumptions or fall back to structural measures (B, W, d_{BE}, r). **Step 3 (objective):** Choose uniform–PAC/regret, instance-dependent BPI, reward-free exploration, offline control/OPE, or partial observability. **Step 4 (verification):** Check realizability/completeness and calibrate confidence via policy certificates when available [4].

Setting	Comp (structure)	Cov (coverage)	**PAC budget** $N(\varepsilon,\delta)$ **(suppressing logs)**
Tabular (online)	$\Phi_{\mathrm{tab}}(S,A,H)$	1	$\tilde{\Theta}\big(\Phi_{\mathrm{tab}}/\varepsilon^2\big)$
Linear MDP (online)	$\mathrm{poly}(d)$	1	$\tilde{\mathcal{O}}\big(\mathrm{poly}(H)\,\mathrm{poly}(d)\,\varepsilon^{-2}\big)$
Kernel/NTK (online)	$d_{\mathrm{eff}}(\lambda)$	1	$\tilde{\mathcal{O}}\big(\mathrm{poly}(H)\,d_{\mathrm{eff}}(\lambda)\,\varepsilon^{-2}\big)$
Bellman rank B (online)	$\mathrm{poly}(B)$	1	$\tilde{\mathcal{O}}\big(\mathrm{poly}(H,B)\,\varepsilon^{-2}\big)$
Witness rank W (model–based)	$\mathrm{poly}(W)$	1	$\tilde{\mathcal{O}}\big(\mathrm{poly}(H,W)\,\varepsilon^{-2}\big)$
Bilinear rank r (online)	$\mathrm{poly}(r)$	1	$\tilde{\mathcal{O}}\big(\mathrm{poly}(H,r)\,\varepsilon^{-2}\big)$
Reward–free (tabular)	Φ_{tab}	exploration budget	$\tilde{\mathcal{O}}\big(S^2A\,\mathrm{poly}(H)\,\varepsilon^{-2}\big)$
Reward–free (linear)	$\mathrm{poly}(d)$	exploration budget	$\tilde{\mathcal{O}}\big(\mathrm{poly}(H)\,\mathrm{poly}(d)\,\varepsilon^{-2}\big)$
Offline (linear/tabular)	$\mathrm{poly}(d)$ or Φ_{tab}	$\mathrm{poly}(C_\star)$	$\tilde{\mathcal{O}}\big(\mathrm{poly}(H)\,\mathrm{Comp}\,\mathrm{poly}(C_\star)\,\varepsilon^{-2}\big)$

Table 4: Rate cookbook in the CSO template. Exact horizon/constant exponents appear in the cited theorems; here we surface the structural capacity and coverage factors that drive $N(\varepsilon,\delta)$.

Meta-insights from CSO factorization.

(i) **Coverage–structure tradeoff:** Rich structure (Comp$\ll SA$) tolerates weaker coverage, but only if data span the structured subspace (e.g., rank-r features must be excited).

(ii) **Objective sets ε-resolution:** Uniform–PAC bounds *all* ε simultaneously; this prevents "epsilon-cheating" and is strictly stronger than single-ε PAC [2].

(iii) **Horizon compounds all factors:** Exponents vary (tabular H^3; linear H^4–H^5; kernel can be higher [35]). Reducing effective horizon (abstraction/discounting/hierarchy) yields superlinear savings.

3 Preliminaries: PAC and RL Foundations

1. Motivation & precise scope. This section fixes notation and formal definitions used throughout the survey and states baseline guarantees that will be referenced repeatedly. The goal is to make explicit: (i) what a Probably Approximately Correct (PAC) guarantee in RL asserts; (ii) how *uniform–PAC* strengthens PAC and connects to high–probability regret; (iii) which distributional and structural assumptions (coverage, realizability, complexity measures) underwrite finite–sample results; and (iv) how offline and reward–free settings are formalized. Clarity at this level is crucial because modern results differ primarily in their assumptions and complexity parameters rather than in proof templates. We adopt episodic finite–horizon MDPs as the default; discounted problems are mapped via an effective horizon $H_{\mathrm{eff}} \asymp (1-\gamma)^{-1}$ when needed. Scope covers definitions and canonical baseline theorems (informal statements with dependencies); sharp constants and refined exponents appear in later sections or in the cited papers. For historical context predating 2018 see Strehl et al. [27], Szepesvári [30], Kaelbling et al. [19], Sutton and Barto [29].

2. Definitions. Let $M = (\mathcal{S},\mathcal{A},P,r,H,\rho)$ be a finite–horizon MDP with $|\mathcal{S}| = S$, $|\mathcal{A}| = A$, horizon $H \in \mathbb{N}$, start distribution ρ, transitions $\{P_h\}_{h=1}^H$, and rewards $r_h \in [0,1]$. A (possibly nonstationary) policy π maps states to distributions over actions. Define V_h^π, Q_h^π in the standard way and $V_1^\pi(\rho) = \mathbb{E}_{s\sim\rho}[V_1^\pi(s)]$.

Definition 1 ((ε,δ)–PAC (fixed–confidence control))**.** Alg *is* (ε,δ)*–PAC if, with probability at least* $1-\delta$*, after at most* $N(\varepsilon,\delta)$ *interactions it outputs* $\hat{\pi}$ *such that* $V_1^\star(\rho) - V_1^{\hat{\pi}}(\rho) \le \varepsilon$*.*

Definition 2 (Uniform–PAC)**.** *With probability at least* $1-\delta$*, for* all $\varepsilon > 0$ *simultaneously the number of* ε*–suboptimal episodes is at most* $N(\varepsilon,\delta)$*. Uniform–PAC implies a high–probability regret bound by summing over* ε *levels [2].*

Definition 3 (Function class; realizability; agnostic)**.** *Let* $\mathcal{F}$ *be a value or* Q*–function class.* Realizability *means* $Q^\star \in \mathcal{F}$ *(or* $V^\star \in \mathcal{F}$*); the* agnostic *setting allows* $Q^\star \notin \mathcal{F}$ *and measures error relative to* $\inf_{f\in\mathcal{F}} \|f - Q^\star\|$*.*

Definition 4 (Covering number)**.** $\mathcal{N}(\epsilon,\mathcal{F},\|\cdot\|)$ *is the size of the smallest* ϵ*–net of* $\mathcal{F}$ *under norm* $\|\cdot\|$*.*

Definition 5 (Bellman operator and error)**.** $(\mathcal{T}_h f)(s,a) = r_h(s,a) + \mathbb{E}_{s'\sim P_h(\cdot|s,a)} \max_{a'} f_{h+1}(s',a')$. *The stagewise Bellman error is* $\|f_h - \mathcal{T}_h f\|$ *in the specified norm.*

Definition 6 (Occupancy and concentrability)**.** *Let* $d_h^\pi(s,a)$ *be the occupancy of policy* π *at stage* h *and* $\mu_h(s,a)$ *the data (behavior) occupancy. Define* $C_\star = \max_h \left\|\frac{d_h^\pi}{\mu_h}\right\|_\infty = \max_{h,(s,a)} \frac{d_h^\pi(s,a)}{\mu_h(s,a)}$.

Definition 7 (Access models)**.** Online*: sequential interaction;* generative model*: i.i.d. sampling from* $P_h(\cdot \mid s,a)$ *on query* (s,a,h)*;* offline*: fixed dataset* $\mathcal{D}$ *with no further interaction.*

Definition 8 (Observation/context space)**.** *In settings with rich observations (§5, §8), let* $\mathcal{X}$ *denote the observation or context space. For standard MDPs,* $\mathcal{X} = \mathcal{S} \times \mathcal{A}$*; for contextual decision processes (CDPs),* $\mathcal{X}$ *may be high-dimensional (e.g., images) while latent states* $z \in [m]$ *with* $m \ll |\mathcal{X}|$ *govern dynamics.*

Definition 9 (Tabular complexity shorthand)**.** *We write* $\Phi_{\text{tab}}(S,A,H) \equiv S\,A\,H^3$ *to denote the canonical tabular factor in episodic sample complexity/regret bounds, up to absolute constants and logarithms.*

3. Canonical results (informal statements with dependencies).

Theorem 5 (Uniform–PAC ⇒ high–probability regret; tabular)**.** *If an algorithm is uniform–PAC with budget* $N(\varepsilon,\delta)$*, then with probability* $1-\delta$ *its cumulative regret after* K *episodes satisfies*

$$\text{Regret}(K) \;=\; \mathcal{O}\Big(\int_0^H \min\{K,\, N(\varepsilon,\delta)\}\, d\varepsilon\Big),$$

which recovers near–minimax tabular regret rates when $N(\varepsilon,\delta)$ *has polynomial dependence in* $(S,A,H,1/\varepsilon)$ *[2].*

Tightness: The conversion is tight; uniform-PAC is strictly stronger than single-ε PAC, and the regret integral cannot be improved in general [2].

Theorem 6 (Tabular minimax sample complexity is $\tilde{\Theta}(SAH^3/\varepsilon^2)$)**.** *In finite-horizon tabular MDPs with* S *states,* A *actions and horizon* H*, the optimal (worst-case)* online *sample complexity to learn an* (ε,δ)*-optimal policy is*

$$N(\varepsilon,\delta) \;=\; \tilde{\Theta}\Big(\frac{SAH^3}{\varepsilon^2}\Big),$$

and the minimax regret after K *episodes is* $\tilde{\Theta}\big(\min\{\sqrt{SAH^3K},\, HK\}\big)$*, both up to polylogarithms in* $(S,A,H,1/\varepsilon,1/\delta)$*. These scalings are achieved by optimistic value-iteration variants and are minimax-sharp up to log factors; see Zhang et al. [36] for tight upper/lower bounds and Domingues et al. [6] for refined lower bounds.*

Tightness: Tight (Θ); matching upper and lower bounds up to logarithms [36, 6].

Theorem 7 (Reward–free exploration (tabular))**.** *There exist algorithms that, without accessing rewards, collect*

$$N_{\text{RFE}}(\varepsilon,\delta) \;=\; \tilde{\mathcal{O}}\Big(\frac{S^2AH^3}{\varepsilon^2}\log\frac{1}{\delta}\Big)$$

episodes and subsequently, for any reward $r \in [0,1]$*, compute an* ε*–optimal policy w.p.* $\geq 1-\delta$*. The* S^2 *factor (vs.* S *in online learning) comes from covering occupancies for* all *downstream rewards [15].*

Tightness: Tight; matching lower bound shows S^2 factor is necessary [15].

Theorem 8 (Linear MDPs; realizable function approximation)**.** *Under linear transition/reward realizability with feature dimension* d*, LSVI–UCB achieves*

$$N(\varepsilon,\delta) \;=\; \tilde{\mathcal{O}}\Big(\frac{d^3H^5}{\varepsilon^2}\log\frac{1}{\delta}\Big) \quad \textit{and} \quad \text{Regret}(K) \;=\; \tilde{\mathcal{O}}\big(d\sqrt{H^3K}\big).$$

Refinements improve the horizon dependence to H^4 *via sharper concentration [?]. Uniform–PAC analogues hold under Bellman completeness [11].*

Tightness: Upper bound only. The H^5 exponent can be reduced to H^4 with variance-aware estimators; optimal H exponent remains open. Dependence on d is believed tight up to polynomial factors.

Theorem 9 (Function classes with bounded Bellman–Eluder dimension)**.** *If the Bellman–Eluder dimension d_{BE} of the value class is finite, sample–efficient learning is possible with rates polynomial in (H, d_{BE}) (precise exponents depend on the algorithmic template) [17].*

Tightness: Upper bound only. Provides general learnability guarantee; lower bounds for specific d_{BE} values are known only in special cases.

Theorem 10 (Offline RL via pessimism; linear MDPs (control))**.** *In offline RL with linear realizability and concentrability $C_\star$, pessimistic algorithms (e.g., PQL) achieve control error ε with dataset size $\tilde{\mathcal{O}}\big(\mathrm{poly}(H)\,\mathrm{poly}(d)\,\mathrm{poly}(C_\star)\,\varepsilon^{-2}\big)$ [26]; in tabular model–based settings, minimax–optimal sample complexity is attained [22].*

Tightness: Upper bound only for linear setting. Tabular model-based offline RL is minimax-tight [22]; linear lower bounds with matching $C_\star$ dependence remain open.

4. Connections to other results.

- **Uniform–PAC as a bridge:** converts fixed–confidence statements into regret bounds and supports per–episode guarantees [2].
- **Coverage as currency:** reward–free exploration *creates* coverage online, while offline pessimism *consumes* coverage quantified by $C_\star$ [15, 26].
- **Structure $\Rightarrow$ learnability:** Bellman rank and Bellman–Eluder dimension delineate when function approximation admits PAC–style guarantees [14, 17].
- **From worst–case to instance–dependent:** identification and gap–based analyses refine tabular rates and motivate adaptive exploration (see Section 4).

5. Practical implications.

- Check assumptions before invoking a bound: realizability ($Q^\star \in \mathcal{F}$), coverage ($C_\star$), and horizon scaling; do not transfer linear–MDP guarantees to arbitrary deep networks.
- For multi–reward or task families, use reward–free exploration to amortize exploration cost across rewards [15].
- In offline pipelines, estimate coverage proxies and prefer pessimistic objectives; avoid deployment when coverage is poor [26].
- When available, compute policy certificates to obtain per–episode accountability [4].

6. Open problems.

i) Uniform–PAC under kernel/over–parameterized classes with minimal spectral assumptions and near–optimal $(H, d_{\mathrm{eff}}, \varepsilon^{-2})$ dependence.

ii) Agnostic PAC guarantees for low–rank/latent models with polynomial dependence on rank/latent size and horizon.

iii) Offline RL under misspecification: bounds separating approximation, estimation, and coverage with sharp trade–offs in $(H, d, C_\star)$.

iv) Data–driven structural selection (SRM for RL): oracle inequalities that adapt to Bellman–Eluder/Bellman rank estimated from trajectories.

7. Related work. **Strehl–Li–Littman (2009)** surveyed PAC–MDP analyses in finite MDPs and set classical $(S, A, H, \varepsilon, \delta)$ baselines [27]. **Szepesvári (2010)** provided a monograph–level treatment of RL algorithms and analysis [30]. **Kaelbling–Littman–Moore (1996)** offered an early broad survey; **Sutton–Barto (2018)** updated foundational algorithmics [19, 29]. **Dann–Lattimore–Brunskill (2017)** introduced uniform–PAC and showed how it yields high–probability regret bounds [2]. **Domingues et al. (2021)** developed tight lower/upper bounds for episodic tabular RL [6]; **Zhang et al. (2024)** gave a sharp characterization of optimal online sample complexity [36]. **Jiang et al. (2017)** formalized Bellman rank and oracle–efficient learning in contextual decision processes [14]; **Jin–Liu–Miryoosefi (2021)** introduced Bellman–Eluder dimension for function approximation [17]. **Jin et al. (2020)** and **Wang et al. (2021)** established learning with linear features and near–optimal regret/PAC dependencies; **He–Zhou–Gu (2021)** derived uniform–PAC for linear MDPs [16, 11]. **Jin et al. (2020)** formalized reward–free exploration with near–optimal dependence on $(S, A, H, 1/\varepsilon)$ [15]. **Shi et al. (2022)** analyzed pessimistic offline RL with concentrability dependence, and **Li et al. (2024)** settled tabular model–based offline complexity [26, 22].

4 Tabular Baselines: Minimax and Instance-Dependent PAC

1. Motivation & precise scope. This section establishes the finite-sample landscape for tabular episodic RL, the setting in which the sharpest PAC and uniform-PAC guarantees are known. The central questions are: (i) what are the *minimax* sample-complexity rates to obtain an ε–optimal policy with confidence $1-\delta$; and (ii) when can *instance-dependent* structure yield substantially faster identification than worst-case rates. The answers calibrate what is statistically achievable without function approximation and serve as baselines for structured models in later sections. We present lower/upper bounds with their dependencies on $(S, A, H, \varepsilon, \delta)$, summarize best-policy identification (BPI) guarantees that depend on suboptimality gaps and reachability, and explain how uniform-PAC connects these fixed-confidence results to regret. We also situate per-episode *policy certificates* that furnish auditable guarantees during learning. Throughout, constants and horizon exponents follow the cited theorems and we defer exact exponents to the cited theorems.

2. Definitions. We work in the finite-horizon tabular MDP $M = (\mathcal{S}, \mathcal{A}, \{P_h\}_{h=1}^H, \{r_h\}_{h=1}^H, H, \rho)$ with $|\mathcal{S}| = S$, $|\mathcal{A}| = A$, rewards $r_h \in [0,1]$, and start distribution ρ. Let V_h^π, Q_h^π be the value and action-value functions of a (possibly nonstationary) policy π, and let $(V^\star, Q^\star)$ denote optimal values.

Definition 10 ((ε,δ)–PAC)**.** *An algorithm is* (ε,δ)–PAC *if, with probability at least* $1-\delta$*, after at most* $N(\varepsilon,\delta)$ *episodes it outputs* $\hat{\pi}$ *such that* $\mathbb{E}_{s\sim\rho}\big[V_1^\star(s) - V_1^{\hat{\pi}}(s)\big] \le \varepsilon$.

Definition 11 (Uniform–PAC)**.** *An algorithm is* uniform–PAC *if, with probability at least* $1-\delta$*, for* all $\varepsilon > 0$ *simultaneously the number of* ε*–suboptimal episodes is at most* $N(\varepsilon,\delta)$*; summing the per-episode suboptimality yields a high-probability regret bound [2].*

Definition 12 (Best-Policy Identification (BPI), gaps, reachability)**.** *For each stage* h *and state* s*, define the* action gap $\Delta_h(s,a) := Q_h^\star(s, a_h^\star(s)) - Q_h^\star(s,a)$ *with* $a_h^\star(s) \in \arg\max_a Q_h^\star(s,a)$*. Let* $d_h^\pi(s)$ *be the state occupancy under* π *at stage* h *and* $q_h(s) := \sup_\pi d_h^\pi(s)$ *a* reachability *factor.* BPI *asks for an algorithm that outputs a policy* $\hat{\pi}$ *with* $\mathbb{E}_{s\sim\rho}[V_1^\star(s) - V_1^{\hat{\pi}}(s)] \le \varepsilon$ *with probability* $1-\delta$ *and whose sample complexity scales with* $\{\Delta_h(s,a), q_h(s)\}$ *rather than* (S, A) *alone.*

Definition 13 (Policy certificate)**.** *A* policy certificate *at episode* t *is a data-dependent upper bound* U_t *such that, with probability at least* $1-\delta$*,* $V_1^\star(\rho) - V_1^{\pi_t}(\rho) \le U_t$*; certificates enable per-episode accountability [4].*

3. Canonical results (informal statements with conditions).

Theorem 11 (Minimax tabular sample complexity)**.** *In finite-horizon tabular MDPs with* S *states,* A *actions, and horizon* H*, the worst-case sample complexity for learning an* (ε,δ)*-optimal policy is*

$$N(\varepsilon,\delta) \;=\; \tilde{\Theta}\Big(\frac{S\,A\,H^3}{\varepsilon^2}\Big),$$

and the minimax regret after K *episodes is* $\tilde{\Theta}\big(\min\{\sqrt{SAH^3K},\, HK\}\big)$*, both up to logarithmic factors. See Domingues et al. [6], Dann et al. [2], Zhang et al. [36] for sharp lower/upper bounds and exact constants.*

Uniform–PAC ⇒ regret. By Theorem 5, uniform–PAC budgets directly imply high-probability regret bounds: integrating $\min\{K, N(\varepsilon,\delta)\}$ over ε recovers near-minimax tabular regret $\tilde{\mathcal{O}}(\sqrt{SAH^3K})$ when $N(\varepsilon,\delta) = \tilde{\mathcal{O}}(SAH^3/\varepsilon^2)$.

Theorem 12 (Instance-dependent identification (BPI))**.** *There exist optimistic identification algorithms whose sample complexity satisfies*

$$N_{\mathrm{BPI}}(\varepsilon,\delta) \;=\; \tilde{\mathcal{O}}\left(\sum_{h=1}^{H}\sum_{s\in\mathcal{S}} q_h(s) \sum_{\substack{a\in\mathcal{A}:\\ \Delta_h(s,a)>\varepsilon}} \frac{1}{\Delta_h(s,a)^2}\right),$$

and lower bounds showing that dependence on inverse squared gaps and reachability is unavoidable up to logarithms [33, 32]. Precise constants and logarithmic factors appear in Wagenmaker et al. [33], Tirinzoni et al. [32].

Theorem 13 (Policy certificates)**.** *For tabular episodic MDPs there exist algorithms that compute per-episode policy certificates U_t such that, with probability at least $1-\delta$, $V_1^\star(\rho) - V_1^{\pi_t}(\rho) \le U_t$ and the cumulative sum of certificates tracks the algorithm's regret up to lower-order terms [4].*

4. Why tabular matters: CSO lens.

- **CSO instantiation:** $\mathsf{Comp} = SAH^3$; $\mathsf{Cov} = 1$ (online); Obj varies (uniform–PAC, BPI, RFE).
- **Calibration role:** Any structured rate (linear, low-rank) must *recover* tabular when structure degenerates (e.g., one-hot features with $d = SA$). Theorem 11 sets the ceiling.
- **Uniform–PAC bridge:** Theorem 5 shows PAC → high-probability regret by integrating counts of ε-bad episodes [2].
- **Instance-dependence:** BPI (Theorem 12) shrinks Comp from SA to a gap- and reachability-weighted sum when $\Delta_h(s,a)$ are large on reachable states.

5. Practical implications.

- Use algorithms with explicit confidence sets (optimism or posterior bonuses) to ensure uniform–PAC behavior; avoid heuristics without quantifiable uncertainty.
- When identification is the goal, prioritize exploration toward states with high estimated reachability and ambiguous gaps; avoid wasting samples where $\Delta_h(s,a)$ is confidently large.
- Track policy certificates during training and gate deployment on $U_t \le \varepsilon$; if certificates stall above threshold, collect more data or revise exploration.
- Treat all PAC claims as conditional on the tabular assumption; do not extrapolate rates to function approximation without verifying realizability and structure.

5b. Mini-example: diagnosing linear realizability. Setting: CartPole with state $(x, \dot{x}, \theta, \dot{\theta})$ and actions $\{\text{left}, \text{right}\}$. Candidate features $\phi(s,a) = [x, \dot{x}, \theta, \dot{\theta}, \mathbf{1}\{a = \text{right}\}]^\top$.

Protocol: (a) collect $n = 500$ random-policy transitions; (b) for each h, fit $\hat{Q}_h = \phi^\top w_h$ via ridge; (c) compute held-out Bellman residuals $\mathcal{E}_h = |\hat{Q}_h - (r_h + \max_{a'} \hat{Q}_{h+1}(s',a'))|$; (d) **pass** if $\frac{1}{n}\sum_i \mathcal{E}_h^2 \lesssim \varepsilon^2$ for target ε. **Failure signal:** residuals grow with h or show bias near $\theta \approx \pm\pi/2$. **Remedy:** add interactions $[\phi, \phi \odot \phi]$ or switch to kernels $\kappa(s,s') = \exp(-\|s-s'\|^2/\sigma^2)$.

Measure	Assumption	Algorithms	Rate prototype	Notes
Bellman rank B	CDP with rank factorization	OLIVE / elimination	$\tilde{\mathcal{O}}(\mathrm{poly}(H, B)\,\varepsilon^{-2})$	Model-free; needs realizability [14].
Witness rank W	Model class & discriminator	Model-based optimism	$\tilde{\mathcal{O}}(\mathrm{poly}(H, W)\,\varepsilon^{-2})$	Witness tests rule out models [28].
BE dim. d_{BE}	Value-class info complexity	BE-optimistic	$\tilde{\mathcal{O}}(\mathrm{poly}(H, d_{\mathrm{BE}})\,\varepsilon^{-2})$	Info-theoretic; general [17].
Linear MDP (d)	Linear realizability/completeness	LSVI-UCB	$\tilde{\mathcal{O}}(\mathrm{poly}(H, d)\,\varepsilon^{-2})$	Elliptical bonuses [16, 11].
Low rank (r)	Rank-r transition factorization	Rep. learning + planning	$\tilde{\mathcal{O}}(\mathrm{poly}(H, r)\,\varepsilon^{-2})$	Subsumes Block MDPs.

Table 5: Which structural knob controls sample complexity? All entries suppress logarithms.

5c. Failure mode: violated Bellman completeness. **Scenario:** 10×10 GridWorld with walls; $\phi(s,a) = [x, y, a]^\top$. Wall "bounce" makes $\mathbb{E}[V_{h+1}(s')|s,a]$ nonlinear. **Symptom:** LSVI-UCB converges suboptimally; residuals plateau ≈ 0.3 for target $\varepsilon = 0.01$; near-wall estimates highly variable. **Root cause:** linear realizability/completeness fails. **Fix:** (i) augment with wall features (adjacent, corner, distance); (ii) use model-based/witness-rank tests; or (iii) accept agnostic error $\varepsilon_{\mathrm{approx}}$ and aim for bounds that degrade gracefully. **Consequence:** Linear guarantees (Theorem 19) do not apply; rates can revert toward tabular $\tilde{\Theta}(SAH^3/\varepsilon^2)$ absent structure.

6. Open problems.

i) Determine the sharp *finite* constants and exact horizon exponents for the minimax PAC sample complexity in episodic tabular MDPs; validate tightness via matching lower bounds (see Zhang et al. [36] for exact exponents)

ii) Design identification algorithms whose instance-dependent rates match lower bounds simultaneously for all gaps $\{\Delta_h(s,a)\}$ and all reachability profiles $\{q_h(s))$ (up to logarithms).

iii) Develop policy-certificate schemes with near-linear-time updates per episode and provable calibration under nonstationary rewards or transitions.

iv) Establish robust uniform–PAC guarantees under small model misspecification in rewards or transitions (e.g., bounded KL or TV perturbations).

7. Related work. **Domingues et al. (2021)** derived refined lower and upper bounds for episodic tabular RL, tightening the worst-case sample complexity landscape [6]. **Zhang et al. (2024)** settled the optimal online sample complexity for tabular RL, clarifying horizon and (S, A) dependences to which PAC upper bounds should align [36]. **Dann, Lattimore, Brunskill (2017)** introduced uniform–PAC and showed how it yields high-probability regret bounds, forging a bridge between PAC and regret analyses [2]. **Wagenmaker, Simchowitz, Jamieson (2022)** established instance-dependent identification rates and corresponding lower bounds, revealing regimes where identification is strictly easier than worst-case learning [33]. **Tirinzoni, Al-Marjani, Kaufmann (2023)** provided optimistic PAC algorithms with instance-dependent guarantees and clarified the role of reachability in BPI [32]. **Dann et al. (2019)** introduced policy certificates enabling per-episode accountability and linked certificate sums to regret [4]. **Jin et al. (2020)** formulated reward-free exploration with near-optimal tabular rates, highlighting the role of exploration-driven coverage for downstream control [15].

5 Structural Complexity for RL

1. Motivation & precise scope. This section formalizes the structural parameters that characterize statistical learnability of reinforcement learning with rich observations and function approximation. The central question is: *which problem-dependent complexity measures control sample complexity and regret beyond tabular MDPs*, and under what assumptions. We focus on four lenses introduced or consolidated during 2017–2021: *Bellman rank* for value-based elimination in contextual decision processes (CDPs), *witness rank* for model-based identification, the *Bellman–Eluder (BE) dimension* as an information-theoretic complexity of value classes, and *bilinear classes* that unify several factorizable settings. These parameters mediate the transition from (S, A)-based tabular rates to *structure-based* rates and expose statistical–computational

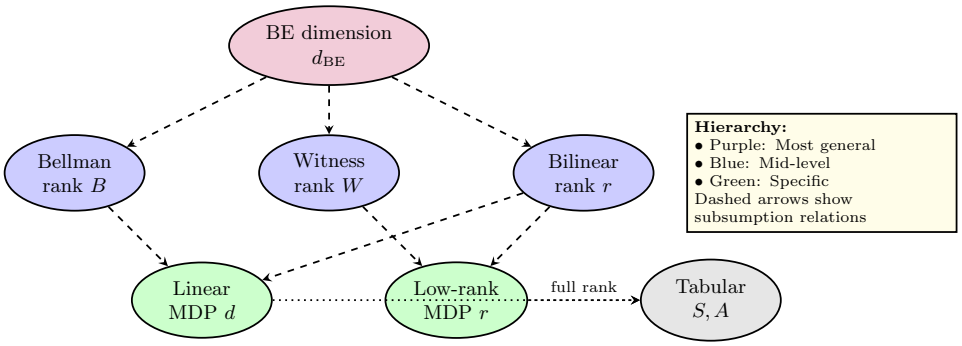

Figure 3: Hierarchy of structural complexity measures. Bellman–Eluder dimension is most general; specific settings are special cases. Tabular RL recovers via one-hot features or full-rank factorizations.

tradeoffs (e.g., oracle efficiency). We state canonical definitions and theorems at a level that makes explicit the dependence on horizon H, accuracy ε, confidence δ, and structural parameters (rank/dimension), deferring constants and log factors to cited sources. Computational assumptions (classification/regression oracles) are stated where needed to avoid NP-hardness barriers.

2. Definitions. Let M be a finite-horizon CDP/MDP with horizon H and a value function class $\mathcal{F} \subseteq \{f = (f_1, \ldots, f_H)\}$, where $f_h : \mathcal{X} \to [0, H-h+1]$ predicts stage-h optimal value proxies; let π_f be a greedy policy w.r.t. f. For a distribution ν_h over contexts/actions at stage h, define the *(expected) Bellman error* of f at stage h by

$$\mathcal{E}_h(f; \nu_h) \;=\; \mathbb{E}_{(x,a)\sim\nu_h}\Big[f_h(x) - r_h(x,a) - \mathbb{E}_{x'\sim P_h(\cdot|x,a)} f_{h+1}(x')\Big].$$

Definition 14 (Bellman rank [14])**.** *A CDP has* Bellman rank B *if for each stage h there exist feature maps $\phi_h : \mathcal{X} \times \mathcal{A} \to \mathbb{R}^B$ and $\psi_h : \mathcal{F} \to \mathbb{R}^B$ with $\|\phi_h\|_2, \|\psi_h\|_2 \le 1$ such that for any $f \in \mathcal{F}$ and any distribution ν_h admissible under policies induced by $\mathcal{F}$, $\mathcal{E}_h(f; \nu_h) = \langle \mathbb{E}_{(x,a)\sim\nu_h} \phi_h(x,a),\, \psi_h(f)\rangle$. The* Bellman rank *is* $\max_h B$.

Definition 15 (Witness rank [28])**.** *Let $\mathcal{M}$ be a model class and $\mathcal{G}$ a discriminator class. For each stage h, suppose there exist feature maps $\Phi_h : \mathcal{X} \times \mathcal{A} \to \mathbb{R}^W$ and $\Upsilon_h : \mathcal{M} \to \mathbb{R}^W$ with bounded norms such that, for any $M, M' \in \mathcal{M}$ and any test $g \in \mathcal{G}$, the* witnessed discrepancy *between M and M' at stage h factors as $\mathbb{E}[g(x_{h+1}) \mid M, (x_h, a_h)] - \mathbb{E}[g(x_{h+1}) \mid M', (x_h, a_h)] = \langle \Phi_h(x_h, a_h),\, \Upsilon_h(M) - \Upsilon_h(M')\rangle$. The smallest such W over admissible choices is the* witness rank.

Definition 16 (Bellman–Eluder (BE) dimension [17])**.** *Given a value class $\mathcal{F}$ and confidence radius $\alpha > 0$, a sequence of queries $(x_1, a_1), \ldots, (x_n, a_n)$ is* BE-independent *if for each t there exist $f, f' \in \mathcal{F}$ such that the (Bellman) predictions for (x_t, a_t) differ by at least α while the cumulative Bellman errors at earlier queries are small. The* BE dimension *$d_{\mathrm{BE}}(\alpha)$ is the maximal length of a BE-independent sequence; write $d_{\mathrm{BE}} = \sup_{\alpha\in(0,1]} d_{\mathrm{BE}}(\alpha)$ if clear from context.*

Definition 17 (Bilinear classes [8])**.** *A sequential decision model is* bilinear of dimension r *if there exist maps $\phi_h : \mathcal{X} \times \mathcal{A} \to \mathbb{R}^r$ and $\psi_h : \mathcal{X} \to \mathbb{R}^r$ such that, for all (x, a) and all h, $\mathbb{E}[f_{h+1}(x') \mid x, a] - \mathbb{E}_{x'\sim\mu_h} f_{h+1}(x') = \langle \phi_h(x,a),\, M_h\, \mathbb{E}_{x'\sim\cdot}[\psi_h(x')]\rangle$ for some stage-dependent matrix M_h with bounded operator norm. (Informally: the Bellman backup residuals admit a rank-r factorization.)*

Definition 18 (Oracle efficiency [3])**.** *An RL algorithm is* oracle-efficient *w.r.t. a function class $\mathcal{F}$ if it runs in time polynomial in $(H, 1/\varepsilon, \log(1/\delta))$ and the cost of calls to standard oracles (e.g., cost-sensitive classification or regression) over $\mathcal{F}$, without enumerating policies or states.*

3. Canonical results (informal statements with conditions).

Theorem 14 (Value-based learning under Bellman rank)**.** *Under realizability and Bellman rank B, elimination-style algorithms (e.g., OLIVE) identify an ε-optimal policy with*

$$N(\varepsilon, \delta) \;=\; \tilde{\mathcal{O}}\big(\mathrm{poly}(B, H)\, \varepsilon^{-2}\, \log(1/\delta)\big),$$

using access to $\mathcal{F}$ for evaluation and greedy improvement [14]. Exact exponents in B and H are given in Jiang et al. [14].

Theorem 15 (Model-based learning under witness rank)**.** *Let $\mathcal{M}$ be a realizable model class with witness rank W relative to a discriminator class $\mathcal{G}$ that separates models. Then a model-based optimistic algorithm achieves PAC control with*

$$N(\varepsilon, \delta) \;=\; \tilde{\mathcal{O}}\big(\mathrm{poly}(W, H)\, \varepsilon^{-2}\, \log(1/\delta)\big),$$

by iteratively ruling out models that are witnessed *inconsistent with data [28].*

Theorem 16 (Learning with finite Bellman–Eluder dimension)**.** *If a value class $\mathcal{F}$ has Bellman–Eluder dimension d_{BE}, there are algorithms with PAC and regret bounds polynomial in (H, d_{BE}) and scaling as ε^{-2}; see Jin et al. [17] for the exact dependencies and constants.*

Theorem 17 (Bilinear classes unify rank-structured RL)**.** *In bilinear classes of dimension r, there exist polynomial-time algorithms with sample complexity*

$$N(\varepsilon, \delta) \;=\; \tilde{\mathcal{O}}\big(\mathrm{poly}(r, H)\, \varepsilon^{-2}\, \log(1/\delta)\big),$$

recovering and unifying guarantees for several previously studied low-rank and feature-based models [8].

Theorem 18 (Oracle-efficient rich-observation RL)**.** *Under realizability and standard reduction oracles over $\mathcal{F}$, there exist oracle-efficient algorithms whose statistical guarantees are governed by structural parameters rather than (S, A); the runtime is polynomial in $(H, 1/\varepsilon, \log(1/\delta))$ and the oracle costs [3].*

4. Connections to other results.

- **Structure $\Rightarrow$ rates:** Bellman rank (B), witness rank (W), BE dimension (d_{BE}), and bilinear dimension (r) replace (S, A) in bounds, yielding problem-dependent rates [14, 28, 17, 8].
- **Model-free vs model-based:** Bellman rank supports hypothesis elimination (model-free), whereas witness rank is tailored to model-based inconsistency testing [28].
- **Bridging to function approximation:** BE dimension provides an information-theoretic complexity of $\mathcal{F}$ that interfaces directly with linear/kernel sections (via effective dimension), complementing linear MDP results [17, 16].
- **Unification by bilinear factorization:** Bilinear classes subsume several structured settings (including certain low-rank and feature-based models) under a common factorization, clarifying when polynomial-time learning is possible [8].
- **Computation via oracles:** Oracle-efficient reductions make the above learnability statements practical for large $\mathcal{F}$ [3].

5. Practical implications.

- Prefer algorithms whose guarantees are expressed in terms of an interpretable structural parameter (e.g., rank or dimension) that can be bounded or estimated from data.
- Use model-based tests when a discriminator class $\mathcal{G}$ is available to *witness* misspecification; otherwise, adopt elimination with explicit Bellman error control.
- Verify realizability surrogates (e.g., Bellman completeness) or deploy robust variants; do not assume tabular rates carry over without structure.
- Leverage oracle-efficient implementations (classification/regression oracles) to keep runtime polynomial while preserving statistical guarantees.

6. Open problems.

i) Tighten the exponents in H and structural parameters $(B, W, d_{\mathrm{BE}}, r)$ to minimax-optimality; supply matching lower bounds for each measure.

ii) Develop data-driven *model selection* that adapts across structures (rank vs BE) with oracle inequalities and uniform-PAC guarantees.

iii) Characterize robustness of structural guarantees under misspecification (e.g., when the true system violates the assumed factorization by a controlled residual).

iv) Design estimators or tests that consistently *bound* B, W, d_{BE}, or r from trajectories with finite-sample confidence.

7. Related work. **Jiang et al. (2017)** introduced *Bellman rank* and the OLIVE algorithm for CDPs, showing PAC sample complexity polynomial in B and H via hypothesis elimination [14]. **Sun et al. (2019)** developed *witness rank* and a model-based optimistic procedure that rules out incorrect models through witnessed moment conditions, yielding PAC rates polynomial in W and H [28]. **Jin, Liu, Miryoosefi (2021)** defined the *Bellman–Eluder dimension*, connecting eluder-style sequential complexity to RL and deriving regret/PAC bounds polynomial in d_{BE} and H [17]. **Du et al. (2021)** formalized *bilinear classes* that factor Bellman residuals and unified several structured models under a rank-r parameter with polynomial sample complexity [8]. **Dann et al. (2018)** established *oracle-efficient* reductions for rich-observation RL, clarifying when statistical guarantees are compatible with polynomial runtime under standard supervised-learning oracles [3].

Computational hardness and oracle efficiency (what rates do not buy you)

Statistical learnability (small B, W, d_{BE}, or r) does not automatically yield polynomial-time algorithms without additional *oracle* assumptions. Bellman-rank style elimination is often *oracle-efficient* (cost-sensitive classification/regression) but can be computationally nontrivial if the oracles are not available or if constraints are nonconvex [3]. Model-based *witness* tests reduce to moment-matching over a discriminator class $\mathcal{G}$; when $\mathcal{G}$ is rich, optimization may be hard. The upshot:

- Prefer *oracle-efficient* reductions; verify that the supervised subproblems are convex and well-posed.
- Treat "polynomial in oracle cost" as a real runtime only when oracles are instantiated (e.g., ridge regression, logistic regression).
- Expect statistical–computational gaps in agnostic settings and in rich CDPs without realizability.

6 Function Approximation: Linear to Kernel/NN

1. Motivation & precise scope. This section studies when reinforcement learning with function approximation admits fixed–confidence (PAC and uniform–PAC) guarantees and how rates depend on structural parameters beyond (S, A). The core question is: *under which realizability or completeness assumptions do linear features, kernel/RKHS models, and over–parameterized neural networks yield sample complexity that scales with an effective dimension rather than the size of the tabular state–action space?* We treat three regimes: (i) *linear MDPs*, where rewards and transition expectations are linear in features; (ii) *kernel/RKHS* models, where value functions lie in a reproducing kernel Hilbert space with an effective dimension; and (iii) *over–parameterized networks* operated in the neural tangent kernel (NTK) regime. We state canonical (informal) results that expose polynomial dependence on horizon H, dimension d or effective dimension $d_{\mathrm{eff}}(\lambda)$, accuracy ε, and confidence δ, and we flag sharp exponents as TODO where needed. We discuss computational assumptions (e.g., least–squares value iteration and linear regression oracles), and we delineate misspecification risks that often surface in practice.

2. Definitions. Let $H \in \mathbb{N}$ be the horizon, rewards $r_h \in [0,1]$, and ρ the start distribution.

Definition 19 (Linear MDP realizability [16])**.** *There exists a known feature map $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ with $\|\phi(s,a)\|_2 \le 1$ and unknown parameters such that: (i) $r_h(s,a) = \phi(s,a)^\top \theta_h^{(r)}$ for some $\theta_h^{(r)} \in \mathbb{R}^d$ with $\|\theta_h^{(r)}\|_2 \le 1$; and (ii) for any bounded measurable $g : \mathcal{S} \to \mathbb{R}$ with $\|g\|_\infty \le 1$ there exists $\theta_{h,g} \in \mathbb{R}^d$, $\|\theta_{h,g}\|_2 \le 1$, satisfying $\mathbb{E}[g(s_{h+1}) \mid s_h = s, a_h = a] = \phi(s,a)^\top \theta_{h,g}$.*

Definition 20 (Bellman completeness for a linear class)**.** *Let $\mathcal{F}_{\text{lin}} = \{f : \exists w_1, \dots, w_H \in \mathbb{R}^d,\ f_h(s,a) = \phi(s,a)^\top w_h\}$.* Bellman completeness *holds if for any $f \in \mathcal{F}_{\text{lin}}$ the Bellman backup $\mathcal{T}_h f$ lies in $\mathcal{F}_{\text{lin}}$ for all h, i.e., $\mathcal{T}_h(\mathcal{F}_{\text{lin}}) \subseteq \mathcal{F}_{\text{lin}}$.*

Definition 21 (Kernel/RKHS value class and effective dimension [35])**.** *Let $(\mathcal{X}, \kappa)$ be a reproducing kernel space with RKHS $\mathcal{H}_\kappa$ and norm $\|\cdot\|_{\mathcal{H}_\kappa}$. Define the (stagewise) value class $\mathcal{F}_\kappa = \{f :\ f_h \in \mathcal{H}_\kappa,\ \|f_h\|_{\mathcal{H}_\kappa} \le B\}$. For a distribution μ over $\mathcal{X}$ with kernel integral operator $\Sigma = \mathbb{E}_{x\sim\mu}[\kappa(x,\cdot) \otimes \kappa(x,\cdot)]$, the* effective dimension *at regularization $\lambda > 0$ is $d_{\text{eff}}(\lambda) := \operatorname{Tr}\big(\Sigma(\Sigma + \lambda I)^{-1}\big)$.*

Definition 22 (NTK regime (over–parameterized networks) [35])**.** *A network trained by (small–step) gradient descent from random initialization is in the* NTK regime *when its predictions evolve approximately linearly in parameters and coincide with kernel regression under the neural tangent kernel κ_{NTK}. Guarantees stated for $\mathcal{F}_\kappa$ transfer to such networks with $\kappa = \kappa_{\text{NTK}}$, up to approximation and optimization errors controlled by width and step size.*

Definition 23 (Generative model vs. online interaction)**.** *A* generative model *allows i.i.d. sampling from $P_h(\cdot \mid s,a)$ upon query (s,a,h). The* online *protocol collects trajectories under the evolving behavior of the algorithm. We use $N(\varepsilon,\delta)$ for sample (episode) complexity and write $\tilde{\mathcal{O}}(\cdot)$ to suppress polylogarithms.*

3. Canonical results (informal statements with conditions).

Theorem 19 (Linear MDPs: optimistic least-squares value iteration)**.** *Under linear MDP realizability with feature dimension d and bounded features, optimistic LSVI achieves PAC sample complexity and high-probability regret with leading terms $N(\varepsilon,\delta) = \tilde{\mathcal{O}}\big(\text{poly}(H,d)\,\varepsilon^{-2}\big)$ and $\text{Regret}(K) = \tilde{\mathcal{O}}\big(\text{poly}(H,d)\sqrt{K}\big)$. Exact exponents and constants appear in Jin et al. [16].*

Theorem 20 (Uniform-PAC for linear MDPs)**.** *If the linear value class is Bellman-complete and realizable, there exist algorithms that are uniform-PAC with budgets $N(\varepsilon,\delta) = \tilde{\mathcal{O}}\big(\text{poly}(H,d)\,\varepsilon^{-2}\big)$, implying high-probability regret via Theorem 5. Precise horizon and dimension exponents are given in He et al. [11].*

Theorem 21 (Kernel/RKHS function approximation)**.** *Let $\mathcal{F}_\kappa$ be an RKHS class with effective dimension $d_{\text{eff}}(\lambda)$. Under the following conditions—(i) realizability ($V^\star \in \mathcal{F}_\kappa$), (ii) Bellman completeness of $\mathcal{F}_\kappa$ under $\mathcal{T}_h$, and (iii) a spectral condition controlling $d_{\text{eff}}(\lambda)$ at regularization $\lambda > 0$—there exist optimistic algorithms with $N(\varepsilon,\delta) = \tilde{\mathcal{O}}\big(\text{poly}(H)\,d_{\text{eff}}(\lambda)\,\varepsilon^{-2}\big)$ and regret $\tilde{\mathcal{O}}\big(\text{poly}(H)\sqrt{d_{\text{eff}}(\lambda)K}\big)$. See Yang et al. [35] for exact dependencies on H, $d_{\text{eff}}(\lambda)$, and λ.*

Theorem 22 (Over–parameterized neural networks via NTK)**.** *For sufficiently wide networks trained in the NTK regime with kernel κ_{NTK} and appropriate step sizes, the guarantees of Theorem 21 apply with $d_{\text{eff}}(\lambda)$ computed for κ_{NTK}, modulo approximation and optimization errors that vanish with width; see Yang et al. [35].*

Theorem 23 (Reward-free exploration under linear approximation)**.** *With linear MDP realizability (dimension d), two-stage reward-free exploration attains $N(\varepsilon,\delta) = \tilde{\mathcal{O}}\big(\text{poly}(H,d)\,\varepsilon^{-2}\big)$ exploratory episodes to guarantee ε-optimality for any downstream reward with probability $1-\delta$; see Wang et al. [34], Kaufmann and Tirinzoni [21] for precise dependencies.*

4. Connections to other results.

- **From tabular to structural rates:** Linear and kernel regimes replace (S,A) by d or $d_{\text{eff}}(\lambda)$, aligning with the structural-complexity perspective of Section 5 and recovering tabular bounds when features span one–hot encodings.

- **Uniform–PAC bridge:** Uniform–PAC results for linear MDPs connect fixed–confidence performance to regret and policy certificates (Sections 1, 4) [11].
- **Kernel/NTK unification:** RKHS analysis subsumes over–parameterized networks in the NTK regime, providing sample complexity in terms of $d_{\text{eff}}(\lambda)$ [35].
- **Coverage creation:** Reward–free exploration amortizes exploration cost across downstream rewards under linear structure (Section 9; 34).
- **Path to offline**: Linear realizability underpins pessimistic offline RL guarantees (Section 10; 26).

5. Practical implications.

- Verify linearity/Bellman completeness with diagnostics (e.g., regression residuals across Bellman backups); if violated, prefer robust or model-based methods rather than invoking linear guarantees.
- In RKHS/NTK pipelines, select regularization λ and norm budgets using validation; monitor $d_{\text{eff}}(\lambda)$ surrogates (e.g., ridge leverage scores) to anticipate sample complexity.
- Use optimism with elliptical bonuses (self-normalized least squares) and maintain well-conditioned design matrices via exploration or perturbations.
- Avoid extrapolating guarantees to arbitrary deep networks outside the NTK regime; when width is modest, treat the model as misspecified.
- In long horizons, track the growth of confidence radii with H; if bounds inflate, redesign features (state abstraction) or shorten effective horizon by discounting.

6. Open problems.

i) Uniform–PAC with kernel/NTK classes under minimal spectral assumptions (no generative model), with rates $\tilde{\mathcal{O}}\big(d_{\text{eff}}(\lambda)\,\text{poly}(H)\,\varepsilon^{-2}\big)$ and polynomial runtime.

ii) Sharp lower bounds for RKHS/NTK RL that match $d_{\text{eff}}(\lambda)$–based upper bounds, including the optimal horizon exponent.

iii) Misspecification-robust function approximation: algorithms whose error decomposes into approximation, estimation, and coverage terms with explicit constants in $(H, d, d_{\text{eff}}(\lambda))$.

iv) Data-driven structure selection between linear and kernel classes with oracle inequalities and uniform–PAC guarantees (SRM for RL).

v) Reward-free exploration with kernel features achieving near–minimax $d_{\text{eff}}(\lambda)$ dependence and computational efficiency.

7. Mini-bibliography (context). **Jin, Yang, Wang (2020)** introduced linear MDP realizability and analyzed optimistic LSVI, yielding polynomial dependence on (H, d) for regret and PAC complexity [16]. **He, Zhou, Gu (2021)** established uniform–PAC bounds for linear MDPs, linking fixed–confidence guarantees to high–probability regret in the linear regime [11]. **Yang et al. (2020)** developed kernel/NTK-based RL analyses, introducing effective dimension $d_{\text{eff}}(\lambda)$ as the capacity term for RKHS and over–parameterized networks [35]. **Wang et al. (2020)** proved reward–free exploration results under linear function approximation, delivering two–stage exploration with polynomial (H, d) dependence [34]. **Dann et al. (2018)** provided oracle-efficient reductions for rich observation RL, clarifying computational assumptions behind LSVI–style methods [3].

7 Deep RL connections: when do guarantees apply?

NTK regime. When networks are sufficiently wide and trained with small steps from random initialization, predictions evolve linearly and coincide with kernel regression under the NTK; our RKHS guarantees (Theorem 21) apply with $d_{\text{eff}}(\lambda)$ computed for κ_{NTK}.

Outside NTK. With moderate width or aggressive optimization, Bellman completeness and realizability may fail. Use the diagnostics in §15.2 and prefer pessimism and OPE in offline settings.

Policy gradient and actor–critic. Finite-sample PAC guarantees are scarce beyond linear/NTK regimes. PAC–Bayes (Section 12) can regularize selection among neural policies via KL-penalized bounds, and certificates can gate deployment.

8 Rich Observations and Low-Rank Structure

Motivation. Many practical problems provide *high-dimensional* observations $x \in \mathcal{X}$ that are generated by, or can be compressed to, a small latent state $z \in [m]$ with $m \ll |\mathcal{X}|$. Two widely studied formalizations are: (i) *Block MDPs*, where each observation x stochastically reveals a *latent* state (decodability); and (ii) *low-rank MDPs*, where transition operators admit a rank-r factorization, enabling representation learning and sample complexity that scales with m or r rather than $|\mathcal{X}|$.

Definitions. **Block MDP** (decodable latent states): there exists a surjective map $\psi : \mathcal{X} \to [m]$ such that $P(x'|x, a)$ depends on x only through $\psi(x)$ and the reward depends on $(\psi(x), a)$. A decoder $\hat{\psi}$ with small error enables *planning in the latent MDP* [7, 23]. **Low-rank MDP**: for each stage h, the transition admits $P_h(s'|s, a) = \langle \phi_h(s, a), \mu_h(s') \rangle$ for bounded features $\phi_h : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^r$, $\mu_h : \mathcal{S} \to \mathbb{R}^r$; the rank r replaces S in rates [1, 5, 12].

Representative guarantees.

Theorem 24 (Block/latent structure: sample complexity)**.** *Under realizability and identifiability (decodability for Block MDPs, or rank-r factorization for low-rank MDPs), there exist polynomial-time algorithms that output an (ε, δ)-optimal policy with*

$$N(\varepsilon, \delta) \;=\; \tilde{\mathcal{O}}\big(\text{poly}(H) \cdot \text{poly}(m) \cdot A\, \varepsilon^{-2}\big) \quad \textit{(Block MDPs)} \quad \textit{or} \quad \tilde{\mathcal{O}}\big(\text{poly}(H) \cdot \text{poly}(r)\, \varepsilon^{-2}\big) \ \textit{(low-rank)},$$

using representation learning and optimistic planning [7, 23, 1, 5, 12].

Connections. (1) These results instantiate the CSO template with $\mathsf{Comp} = m$ (latent size) or r (rank). (2) Witness/Bellman ranks from §5 often remain small when a low-dimensional latent structure exists, explaining polynomial dependence on m or r. (3) Block/low-rank structure complements linear/kernalized value classes (§6); linear features may recover the same latent geometry.

Practical implications. *When the observation is high-dimensional but compressible*, look for diagnostics of decodability/low-rankness: contrastive state decoding, spectral footprints (empirical rank), and success of simple linear Q-heads. Prefer algorithms that (a) learn a decoder or (b) fit a factored model before value iteration. Use per-episode certificates when available to gate deployment.

Related work. Du et al. [7] initiated provably efficient RL with rich observations via latent-state decoding (Block MDPs). Misra et al. [23] (HOMER) gave a decodable representation learning approach with exploration. Agarwal et al. [1] analyzed model-based exploration with learned low-rank features (FLAMBE). Dann et al. [5] studied agnostic low-rank RL; Huang et al. [12] proposed density-feature methods.

9 Reward-Free Exploration

Motivation. In many applications the reward is not specified during data collection (e.g., multi-task evaluation, safety audits). *Reward-free exploration* (RFE) asks the learner to collect data so that, *for any* reward $r \in [0,1]$ specified later, we can compute an ε-optimal policy with confidence $1-\delta$ without further interaction.

Template. Two-stage procedures (i) explore to build a model or confidence sets that cover the dynamics and state-action occupancy, then (ii) when a reward is revealed, plan with the learned model/sets. Coverage is the currency: the first stage must create enough support for *all* plausible rewards.

Representative guarantees.

Theorem 25 (Tabular RFE)**.** *There exist algorithms that, without accessing rewards, collect* $\tilde{\mathcal{O}}\big(S^2 A \,\mathrm{poly}(H)\, \varepsilon^{-2}\big)$ *episodes and subsequently, for any reward* $r \in [0,1]$*, compute an* ε*-optimal policy with probability* $1-\delta$ *[15].*

Theorem 26 (Linear function approximation)**.** *Under linear MDP realizability with feature dimension* d*, two-stage RFE achieves* $\tilde{\mathcal{O}}\big(\mathrm{poly}(H)\,\mathrm{poly}(d)\,\varepsilon^{-2}\big)$ *samples to guarantee* ε*-optimality for any reward, again with probability* $1-\delta$ *[34, 21].*

Connections and practice. RFE operationalizes the CSO template where the "coverage" knob is an *exploration budget* created up front. It is particularly effective in multi-reward/multi-task evaluation pipelines. In long horizons, design exploration policies that keep design matrices well-conditioned (elliptical bonuses or randomized probing).

Related work. Jin et al. [15] formalized RFE and gave near-optimal tabular rates. Wang et al. [34] extended RFE under linear function approximation, and Kaufmann and Tirinzoni [21] developed adaptive strategies that reduce sample sizes in benign instances.

10 Offline RL: Pessimism, OPE, and Model-Based Learning

Motivation. In offline RL we receive a fixed dataset $\mathcal{D}$ collected by an unknown behavior policy and must learn without further interaction. The key challenge is *coverage*: does $\mathcal{D}$ support the target policy? Pessimism (lower confidence bounds) protects against over-optimism outside support.

Pessimism and coverage. Let $C_\star = \max_h \left\| \frac{d_h^\pi}{\mu_h} \right\|_\infty$ be a (clipped) concentrability coefficient (Definition 6). Pessimistic value iteration (PEVI) and pessimistic Q-learning (PQL) control error by subtracting uncertainty penalties calibrated to coverage.

Theorem 27 (Linear MDPs: pessimistic control)**.** *Under linear MDP realizability with feature dimension* d *and concentrability* $C_\star$*, pessimistic algorithms (PEVI/PQL) achieve control error* ε *with dataset size*

$$n \;=\; \tilde{\mathcal{O}}\Big(\mathrm{poly}(H)\,\mathrm{poly}(d)\,\mathrm{poly}(C_\star)\,\varepsilon^{-2}\Big),$$

with polynomial-time computation [18, 26].

Model-based offline RL (tabular). Model-based estimators with uncertainty-aware planning are minimax-optimal in tabular settings:

Theorem 28 (Tabular offline sample complexity (model-based, minimax))**.** *There exist model-based offline procedures whose sample complexity to achieve* ε*-optimal control is minimax-optimal (up to logs) with explicit dependence on* (S, A, H) *and coverage; see Li et al. [22] for a sharp characterization.*

Off-policy evaluation (OPE). For policy evaluation, doubly-robust and semiparametric efficient estimators attain optimal asymptotic variance under standard conditions:

Theorem 29 (Efficient OPE)**.** *Under mild regularity, there exist OPE estimators that achieve statistical efficiency, and minimax value intervals can be constructed with finite-sample validity [20, 13].*

Practice and pitfalls. Estimate coverage proxies (e.g., density ratios, ridge leverage scores) and *prefer pessimism* when proxies indicate partial support. Avoid policy improvement if covariate shift is severe; use OPE to gate deployment. If realizability is doubtful, prefer model-based or robust objectives.

CSO lens (offline). $\mathsf{Comp} = \mathrm{poly}(d)$ (linear) or $\Phi_{\mathrm{tab}}(S, A, H)$ (tabular model-based); $\mathsf{Cov} = \mathrm{poly}(C_\star)$ from the dataset (coverage is the bottleneck); Obj = pessimistic control or OPE. **Diagnosis:** if a proxy $\widehat{C}_\star$ is large, prefer OPE/abstain; if small, use PEVI/PQL [18, 26].

Related work. Jin et al. [18] introduced PEVI; Shi et al. [26] analyzed PQL with explicit $C_\star$ dependence. Li et al. [22] settled model-based tabular offline sample complexity. Kallus and Uehara [20] derived statistically efficient OPE; Jiang and Huang [13] gave minimax value intervals.

11 Partial Observability and Confounding

Identifiable latent-state POMDPs and Block MDPs. When observations admit a decodable latent representation with small m, sample-efficient learning is possible by learning a decoder and planning in latent space (cf. §8). Under separation/mixing and identifiability, algorithms such as HOMER achieve sample complexity polynomial in $(m, H, 1/\varepsilon)$ [23].

Low-rank observability. Beyond exact decoding, low-rank transition factorizations let us exploit representation learning without a strict $\psi : \mathcal{X} \to [m]$ mapping (Theorem 24).

Confounded OPE. In observational datasets, unobserved confounders bias OPE. Structure (e.g., proxies, IV-style assumptions) and sensitivity analyses yield valid intervals; see Shi et al. [25]. *Practice:* prefer interval estimates; avoid point improvement unless sensitivity bounds are tight.

Takeaways. (i) Partial observability is tractable under identifiability/low-rank structure. (ii) For confounded OPE, deploy intervals and abstain when sensitivity is high.

12 PAC–Bayes for RL

Motivation. PAC–Bayes bounds give *data-dependent*, distributional generalization guarantees for randomized predictors or policies, balancing empirical performance with a complexity term (KL divergence to a prior).

Generic bound (evaluation). Let Π be a policy class and let $\hat{R}(\pi)$ be an empirical risk (e.g., negative return estimator) computed from n i.i.d. trajectories. For any prior P on Π, with probability at least $1 - \delta$, for all posteriors Q on Π,

$$\mathbb{E}_{\pi\sim Q}\big[R(\pi)\big] \;\le\; \mathbb{E}_{\pi\sim Q}\big[\hat{R}(\pi)\big] \;+\; \sqrt{\frac{\mathrm{KL}(Q\|P) + \ln(1/\delta)}{2n}} \;+\; \text{bias/variance terms from RL estimation,}$$

leading to high-probability guarantees for policy *evaluation* and *selection* [9, 24, 10, 31].

Connections and practice. PAC–Bayes complements pessimism in offline RL: one can treat the evaluation error as the empirical term and add a KL-penalty to regularize policy selection. In continuous control, priors can be induced by behavior policies or model classes.

Related work. Fard et al. [9] established PAC–Bayes for policy evaluation. Rivasplata et al. [24] handled unbounded losses; Flynn et al. [10] surveyed PAC–Bayes for bandits.

13 Synthesis: What is Settled vs. What is Open

Settled (up to logs/constant factors).

- **Tabular online RL:** near-tight minimax sample complexity/regret and the uniform-PAC $\leftrightarrow$ regret bridge [6, 36, 2].
- **Reward-free (tabular):** near-optimal dependence on $(S, A, H, 1/\varepsilon)$ [15].
- **Linear MDPs:** polynomial dependence on (H, d) for PAC/regret; uniform-PAC extensions [16, 11].
- **Offline (tabular model-based):** minimax sample complexity settled [22]; pessimistic linear-MDP control with explicit $C_\star$ dependence [26].

Open (precise targets).

- **Kernel/NTK:** uniform-PAC with optimal $d_{\text{eff}}(\lambda)$ and horizon exponents under online access and without Bellman completeness.
- **Agnostic low-rank/latent:** PAC rates with polynomial dependence on rank/latent size and H, robust to mild misspecification.
- **Offline under misspecification:** sharp decompositions of error into approximation, estimation, and coverage with tractable algorithms.
- **Instance-dependent FA:** gap-based identification beyond tabular (linear/kernel/low-rank), including lower bounds.
- **Data-driven structure selection:** SRM-style model selection across linear/kernel/low-rank with oracle inequalities and uniform-PAC.

CSO retrospective. Through the CSO lens ($N \approx \mathsf{Cov} \times \mathsf{Comp} \times \text{poly}(H) \times \varepsilon^{-2}$): *Coverage* is settled for online/RFE and open for partial offline coverage estimation and adaptation; *Structure* is clear for linear/low-rank and open for kernel/NTK and agnostic models; *Objective* variations are settled in tabular, partly open with function approximation (instance-dependent FA). The hardest frontiers couple knobs (coverage + structure; structure + misspecification).

How to use the taxonomy. Locate your problem in Table 2 by *access* and *structure*, pick the *objective*, and read off the best available rate and algorithm class; verify assumptions and gate with certificates before deployment.

14 Open Problems: Within Reach vs. Frontier

Within 12–24 months (Within Reach)

W1 Uniform–PAC with kernels under practical completeness. Target: $N = \tilde{O}(d_{\text{eff}}(\lambda)\,\text{poly}(H)\,\varepsilon^{-2})$ under online access, with completeness verified by data-driven residual tests rather than assumed.

W2 Instance-dependent identification with FA. Gap-based BPI for linear/low-rank models matching tabular-style lower bounds (inverse squared gaps weighted by reachability).

W3 Scalable coverage estimation with guarantees. Ridge leverage and density-ratio estimators with finite-sample CIs to gate offline policy improvement.

3–5 years (Frontier)

F1 Agnostic low-rank/latent PAC. Rates polynomial in (r, H) robust to misspecification $\varepsilon_{\text{approx}}$ with oracle-efficient algorithms.

F2 Offline RL under misspecification. Sharp decomposition $\varepsilon_{\text{tot}} = \varepsilon_{\text{approx}} + \varepsilon_{\text{est}} + \varepsilon_{\text{cov}}$ with tractable pessimistic algorithms and tight constants.

F3 Lower bounds for RKHS/NTK RL. Information-theoretic lower bounds that match $d_{\text{eff}}(\lambda)$-based rates and clarify optimal horizon exponents and completeness necessity.

Impossibility and necessity frontiers

- **Offline without coverage:** Without assumptions bounding $C_\star$ (or similar), consistent improvement over μ is impossible; OPE intervals are the right object.
- **Kernel RL without completeness:** In general CDPs, RKHS value classes that are not closed under Bellman updates can exhibit *aliasing* that prevents PAC guarantees without additional structure or a generative model; completeness (or verified residual smallness) is necessary.
- **Agnostic FA:** With unrestricted function classes and no realizability, polynomial sample complexity with polynomial time is unlikely; oracle assumptions or structure are needed.

15 Practical Toolkit: Coverage, Diagnostics, and Deployment

15.1 Offline coverage estimation: concrete procedures

Goal. Decide whether a fixed dataset $\mathcal{D}$ adequately supports policy improvement. We summarize three complementary diagnostics.

1. **Density-ratio (behavior vs. candidate):** For each stage h, fit $w_h(x,a) = \frac{d_h^\pi(x,a)}{\mu_h(x,a)}$ via *logistic regression* on labeled pairs from (i) a rollout buffer of π under a learned model or doubly-robust simulator, and (ii) samples from $\mathcal{D}$ (negative class). Use clipping at $w_{\max}$ and report:

$$\text{ESS}_h = \frac{(\sum_i w_{h,i})^2}{\sum_i w_{h,i}^2}, \quad w_{h,0.99}, \quad \max_i w_{h,i}.$$

   Gate: if $\min_h \text{ESS}_h \ll 100$ or $w_{h,0.99} \gg 50$, abstain or shrink π toward μ.

2. **Ridge leverage scores (feature coverage):** For linear / kernel value estimation, compute stagewise design matrices X_h; let $\tau_{h,i} = x_{h,i}^\top (X_h^\top X_h + \lambda I)^{-1} x_{h,i}$. Report $\bar{\tau}_h$ and upper quantiles. **Gate:** large mass of τ near 1 with poor dispersion indicates weak coverage.

3. **State-action sparsity maps:** Bin (x,a) (or embeddings) and visualize occupancy heatmaps across h. **Gate:** large dark regions where π visits but μ does not $\Rightarrow$ pessimism or abstention.

Algorithm 1 CoverageGate$(\mathcal{D}, \pi, \lambda, w_{\max})$

```
for h = 1 ... H do
    Fit logistic reg. classifier to distinguish (x_h, a_h) ~ D vs. (x_h, a_h) ~ rollouts of π under a learned model; obtain ŵ_h.
    Clip: ŵ_h ← min(ŵ_h, w_max). Compute ESS_h, q_0.99(w_h), max w_h.
    Compute ridge leverage scores using X_h and λ; summarize τ̄_h, q_0.95(τ_h).
end for
return PASS if min_h ESS_h ≥ 200 and q_0.99(w_h) ≤ 30 and q_0.95(τ_h) ≤ 0.5; else WARN/ABSTAIN.
```

Implementation notes. Use cross-fitting for nuisance models (behavior policy, dynamics) to reduce bias. For kernels, approximate leverage via Nyström features.

15.2 Misspecification diagnostics for linear/Kernel

Algorithm 2 BellmanResidualTest($\phi, \mathsf{Alg}, \varepsilon$)

1: Collect n random-policy episodes; fit $f_h(s,a) = \phi(s,a)^\top w_h$ by ridge.
2: Compute held-out residuals $\mathcal{E}_h = |f_h - (r_h + \max_{a'} f_{h+1}(s',a'))|$.
3: **Pass** if $\frac{1}{n}\sum_i \mathcal{E}_h^2 \le \varepsilon^2$ for all h; otherwise enrich features or move to kernel/low-rank models.

15.3 Quantitative deployment gates

We recommend a two-gate rule:

- **Coverage gate (Alg. 1)**: $\min_h \mathrm{ESS}_h \ge 200$, $q_{0.99}(w_h) \le 30$, $q_{0.95}(\tau_h) \le 0.5$.
- **Certificate gate**: deploy only if per-episode policy certificate $U_t \le \varepsilon$ (Theorem 13); else collect data or shrink π.

15.4 Worked numerical sketch

On a 10×10 gridworld ($H = 30$), random behavior μ and target π trained offline: $\min_h \mathrm{ESS}_h = 74$, $q_{0.99}(w_h) = 115$, $q_{0.95}(\tau_h) = 0.81 \Rightarrow$ Abstain; after 200 new online episodes guided by pessimism: $\min_h \mathrm{ESS}_h = 312$, $q_{0.99}(w_h) = 21$, $q_{0.95}(\tau_h) = 0.42 \Rightarrow$ Pass.

16 Ethics Statement

This survey aggregates theory rather than proposing new algorithms, but its claims can influence practice in safety–critical domains (e.g., healthcare, robotics, education). The primary ethical risk is *misapplication of guarantees outside their assumptions*—for example, invoking linear–MDP bounds with misspecified features, or deploying an offline policy under poor coverage, which can harm users. Mitigations emphasized throughout include: (i) explicit assumption checklists (realizability, horizon scaling, coverage proxies); (ii) per–episode *policy certificates* to gate deployment; (iii) abstention and interval–valued *off–policy evaluation* (OPE) when coverage is weak; and (iv) reporting structural and coverage diagnostics alongside results. Fairness risks arise when offline datasets encode historical bias; coverage analysis and sensitivity checks can reveal regions where performance is unreliable. We recommend conservative defaults (pessimism, OPE before improvement), public documentation of assumptions, and auditing procedures that include the failure modes cataloged in this survey.

Reproducibility Statement

- **Scope & inclusion criteria.** Time window 2018–2025; we include peer–reviewed results (JMLR; PMLR venues COLT/ALT/ICML; NeurIPS) and widely cited arXiv preprints when anchoring common bounds; pre–2018 works are cited for foundations.
- **Problem models & notation.** Unified MDP/POMDP/CDP notation in §3, a consolidated glossary (App. A), and first–use definitions for $H, S, A, d, r, B, W, d_{\mathrm{BE}}, d_{\mathrm{eff}}(\lambda), C_\star$.
- **What is reproduced.** Canonical guarantees are restated with explicit parameter dependencies; rate tables flag where logarithms are suppressed; primary sources are cited for constants and exact exponents.

- **Search protocol (summary).** Queries: *uniform–PAC reinforcement learning*, *PAC RL finite horizon*, *Bellman rank*, *Bellman–Eluder dimension*, *linear MDP*, *low–rank MDP / Block MDP*, *reward–free exploration*, *offline RL pessimism / concentrability*, *PAC–Bayes reinforcement learning*. Venues scanned: JMLR; PMLR (COLT/ALT/ICML); NeurIPS; selected arXiv for works later published. We favored papers with clear assumptions and finite–sample theorems.
- **Data/code.** No new datasets or code are introduced; figures are conceptual. When algorithms are discussed, we reference canonical implementations where applicable.
- **Disambiguation.** When rates differ across papers (e.g., horizon exponents), we state the dependency class and cite the precise theorem for constants; competing viewpoints are mentioned where relevant.
- **Versioning.** Citations correspond to versions current as of October 2025.

Limitations

Our focus is episodic finite–horizon and fixed–confidence (PAC / uniform–PAC) guarantees; we do not cover continuous–time control, average–reward/ergodic settings, or detailed deep–RL practice beyond linear/kernel/NTK regimes. Several summaries suppress polylogarithms; constants and sharp horizon exponents are deferred to primary sources. Our treatment of partial observability is limited to identifiable latent–state / low–rank subclasses and confounded OPE. Many results assume realizability or Bellman completeness; robustness to misspecification is highlighted as an open problem. Finally, while we provide diagnostics and decision rules, we do not release audited toolchains for coverage estimation or certificate computation.

Bibliography organization. References are grouped thematically: *Foundations* (Strehl, Szepesvári, Kaelbling, Sutton–Barto); *Uniform–PAC & tabular* (Dann–Lattimore–Brunskill; Domingues; Zhang); *Structural complexity* (Jiang; Sun; Jin–Liu–Miryoosefi; Du–Kakade); *Function approximation* (Jin–Yang–Wang; He–Zhou–Gu; Yang et al.); *Rich observations* (Du; Misra; Agarwal; Dann–agnostic;); *Reward-free* (Jin; Kaufmann; Wang); *Offline* (Jin–PEVI; Shi–PQL; Li–AOS; Kallus; Jiang); *PAC–Bayes* (Fard; Rivasplata; Flynn; Tasdighi).

References

[1] Alekh Agarwal, Sham Kakade, Akshay Krishnamurthy, and Wen Sun. Flambe: Structural complexity and representation learning of low rank MDPs. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, 2020. URL `https://proceedings.neurips.cc/paper/2020/hash/e894d787e2fd6c133af47140aa156f00-Abstract.html`. arXiv:2006.10814.

[2] Christoph Dann, Tor Lattimore, and Emma Brunskill. Unifying PAC and regret: Uniform-PAC bounds for episodic RL. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. URL `https://arxiv.org/abs/1703.07710`.

[3] Christoph Dann, Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E. Schapire. On oracle-efficient PAC RL with rich observations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. URL `https://papers.nips.cc/paper/7416-on-oracle-efficient-pac-rl-with-rich-observations`.

[4] Christoph Dann, Lihong Li, Wei Wei, and Emma Brunskill. Policy certificates: Towards accountable reinforcement learning. In *International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*, pages 1507–1516, 2019. URL `https://proceedings.mlr.press/v97/dann19a.html`.

[5] Christoph Dann, Yishay Mansour, Mehryar Mohri, Ayush Sekhari, and Karthik Sridharan. Agnostic reinforcement learning with low-rank MDPs and rich observations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. URL `https://proceedings.neurips.cc/paper/2021/hash/9eed867b73ab1eab60583c9d4a789b1b-Abstract.html`. arXiv:2106.11519.

[6] Omar Darwiche Domingues, Pierre Ménard, Emilie Kaufmann, and Michal Valko. Episodic reinforcement learning in finite MDPs: Minimax lower bounds revisited. In Vitaly Feldman, Katrina Ligett, and Sivan Sabato, editors, *Proceedings of the 32nd International Conference on Algorithmic Learning Theory*, volume 132 of *Proceedings of Machine Learning Research*, pages 578–598. PMLR, 2021. URL https://proceedings.mlr.press/v132/domingues21a.html.

[7] Simon S. Du, Akshay Krishnamurthy, Nan Jiang, Alekh Agarwal, Miroslav Dudík, and John Langford. Provably efficient RL with rich observations via latent state decoding. In *International Conference on Machine Learning (ICML)*, volume 97 of *PMLR*, pages 1665–1674, 2019. URL https://proceedings.mlr.press/v97/du19b.html.

[8] Simon S. Du, Sham M. Kakade, Jason D. Lee, Shachar Lovett, Gaurav Mahajan, Wen Sun, and Ruosong Wang. Bilinear classes: A structural framework for provable generalization in RL. In *International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pages 2826–2836, 2021. URL https://proceedings.mlr.press/v139/du21a.html.

[9] Mahdi Milani Fard, Joelle Pineau, and Csaba Szepesvári. PAC-bayesian policy evaluation. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 22 of *PMLR*, 2012. URL https://proceedings.mlr.press/v22/fard12.html.

[10] Hamish Flynn, David Reeb, Melih Kandemir, and Jan Peters. Pac-bayes bounds for bandit problems: A survey and experimental comparison. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. doi: 10.1109/TPAMI.2023.3305381. URL https://dl.acm.org/doi/10.1109/TPAMI.2023.3305381. See also arXiv:2211.16110.

[11] Jiafan He, Dongruo Zhou, and Quanquan Gu. Uniform-PAC bounds for reinforcement learning with linear function approximation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/7695ea769f021803c508817dd374bb27-Abstract.html. Also available as arXiv:2106.11612.

[12] Audrey Huang, Jinglin Chen, and Nan Jiang. Reinforcement learning in low-rank MDPs with density features. In *International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pages 13710–13752, 2023. URL https://proceedings.mlr.press/v202/huang23a.html. arXiv:2302.02252.

[13] Nan Jiang and Jiawei Huang. Minimax value interval for off-policy evaluation and policy optimization. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, pages 2747–2758, 2020. URL https://papers.nips.cc/paper/2020/hash/1cd138d0499a68f4bb72bee04bbec2d7-Abstract.html. arXiv:2002.02081.

[14] Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E. Schapire. Contextual decision processes with low bellman rank are PAC-learnable. In *International Conference on Machine Learning (ICML)*, volume 70 of *PMLR*, 2017. URL https://proceedings.mlr.press/v70/jiang17c.html.

[15] Chi Jin, Akshay Krishnamurthy, Max Simchowitz, and Tengyu Yu. Reward-free exploration for reinforcement learning. In *International Conference on Machine Learning (ICML)*, volume 119 of *Proceedings of Machine Learning Research*, pages 4870–4879, 2020. URL https://proceedings.mlr.press/v119/jin20d.html.

[16] Chi Jin, Zhuoran Yang, and Zhaoran Wang. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory (COLT)*, volume 125 of *PMLR*, 2020. URL https://proceedings.mlr.press/v125/jin20a.html.

[17] Chi Jin, Qinghua Liu, and Sobhan Miryoosefi. Bellman-eluder dimension: New rich classes of RL problems, and sample-efficient algorithms. In *Advances in Neural Information Processing Systems*, volume 34, pages 13406–13418, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/6f5e4e86a87220e5d361ad82f1ebc335-Abstract.html.

[18] Ying Jin, Zhuoran Yang, and Zhaoran Wang. Is pessimism provably efficient for offline RL? In *International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, 2021. URL https://proceedings.mlr.press/v139/jin21e.html.

[19] Leslie Pack Kaelbling, Michael L. Littman, and Andrew W. Moore. Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*, 4:237–285, 1996.

[20] Nathan Kallus and Masatoshi Uehara. Statistically efficient off-policy evaluation for reinforcement learning. *Journal of Machine Learning Research*, 21(167):1–63, 2020. URL http://jmlr.org/papers/v21/20-346.html.

[21] Emilie Kaufmann and Andrea Tirinzoni. Adaptive reward-free exploration. In *Algorithmic Learning Theory (ALT)*, volume 132 of *PMLR*, 2021. URL https://proceedings.mlr.press/v132/tirinzoni21a.html.

[22] Gen Li, Laixi Shi, Yuxin Chen, Yuejie Chi, and Yuting Wei. Settling the sample complexity of model-based offline reinforcement learning. *Annals of Statistics*, 2024. URL https://projecteuclid.org/journals/annals-of-statistics/volume-52/issue-1/Settling-the-sample-complexity-of-model-based-offline-reinforcement-learning/10.1214/23-AOS2342.pdf.

[23] Dipendra Misra, Mikael Henaff, Akshay Krishnamurthy, and John Langford. Kinematic state abstraction and provably efficient rich-observation reinforcement learning. In *International Conference on Machine Learning (ICML)*, volume 119 of *Proceedings of Machine Learning Research*, pages 6961–6971, 2020. URL https://proceedings.mlr.press/v119/misra20a.html.

[24] Omar Rivasplata, Ilja Kuzborskij, Csaba Szepesvári, and John Shawe-Taylor. Pac-bayes analysis beyond the usual bounds. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/c3992e9a68c5ae12bd18488bc579b30d-Abstract.html. arXiv:2006.13057.

[25] Jikai Shi, Dongruo Zhou, and Quanquan Gu. Off-policy evaluation in confounded POMDPs. In *International Conference on Machine Learning (ICML)*, volume 162 of *PMLR*, 2022. URL https://proceedings.mlr.press/v162/shi22a.html.

[26] Laixi Shi, Gen Li, Yuting Wei, Yuxin Chen, and Yuejie Chi. Pessimistic q-learning for offline reinforcement learning: Towards optimal sample complexity. In *International Conference on Machine Learning (ICML)*, volume 162 of *Proceedings of Machine Learning Research*, pages 19967–20025, 2022. URL https://proceedings.mlr.press/v162/shi22b.html.

[27] Alexander L. Strehl, Lihong Li, and Michael L. Littman. PAC analysis of a class of reinforcement learning algorithms. *Journal of Machine Learning Research*, 10:2413–2444, 2009. URL http://jmlr.org/papers/v10/strehl09a.html.

[28] Wen Sun, Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, and John Langford. Model-based reinforcement learning in contextual decision processes: PAC bounds and exponential improvements over model-free approaches. In *Conference on Learning Theory (COLT)*, volume 99 of *Proceedings of Machine Learning Research*, pages 2898–2933, 2019. URL https://proceedings.mlr.press/v99/sun19a.html.

[29] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction (2nd ed.)*. MIT Press, 2018.

[30] Csaba Szepesvári. *Algorithms for Reinforcement Learning*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool, 2010.

[31] Bahareh Tasdighi, Manuel Haussmann, Nicklas Werge, Yi-Shan Wu, and Melih Kandemir. Deep exploration with PAC-bayes. *arXiv preprint*, 2024. URL https://arxiv.org/abs/2402.03055.

[32] Andrea Tirinzoni, Mohammed Al-Marjani, and Emilie Kaufmann. Optimistic PAC reinforcement learning: the instance-dependent view. In *Algorithmic Learning Theory (ALT)*, volume 201 of *PMLR*, 2023. URL `https://proceedings.mlr.press/v201/tirinzoni23a.html`.

[33] Andrew J. Wagenmaker, Max Simchowitz, and Kevin Jamieson. Beyond no-regret: Instance-dependent PAC reinforcement learning. In *Conference on Learning Theory (COLT)*, volume 178 of *PMLR*, 2022. URL `https://proceedings.mlr.press/v178/wagenmaker22a.html`.

[34] Ruosong Wang, Simon S. Du, Lin F. Yang, and Ruslan Salakhutdinov. On reward-free reinforcement learning with linear function approximation. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, 2020. URL `https://proceedings.neurips.cc/paper/2020/hash/ce4449660c6523b377b22a1dc2da5556-Abstract.html`. arXiv:2006.11274.

[35] Zhuoran Yang, Chi Jin, Zhaoran Wang, Mengdi Wang, and Michael I. Jordan. Provably efficient reinforcement learning with kernel and neural function approximations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. URL `https://proceedings.neurips.cc/paper/2020/hash/9fa04f87c9138de23e92582b4ce549ec-Abstract.html`.

[36] Zihan Zhang, Yuxin Chen, Jason D Lee, and Simon S Du. Settling the sample complexity of online reinforcement learning. In Shipra Agrawal and Aaron Roth, editors, *Proceedings of Thirty Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pages 5213–5219. PMLR, 30 Jun–03 Jul 2024. URL `https://proceedings.mlr.press/v247/zhang24a.html`.

A Notation and Abbreviations

This appendix consolidates all notation in consolidated tables for quick reference.

Core MDP Notation

Symbol	Meaning	First Defined
$\mathcal{S}, \mathcal{A}, \mathcal{X}$	State, action, observation/context spaces	Preliminaries
$S = \lvert\mathcal{S}\rvert$, $A = \lvert\mathcal{A}\rvert$	Cardinalities (finite case)	Preliminaries
H	Horizon (episode length in timesteps)	Preliminaries
γ	Discount factor; $H_{\text{eff}} \approx (1-\gamma)^{-1}$	Preliminaries
ρ	Start-state distribution; $s_0 \sim \rho$	Preliminaries
P_h, r_h	Stage-h transition kernel and reward	Preliminaries
π	Policy (possibly nonstationary)	Preliminaries
V_h^π, Q_h^π	Value and action-value functions	Preliminaries
$\pi^\star, V^\star, Q^\star$	Optimal policy and values	Preliminaries

Sample Complexity and Confidence

Symbol	Meaning	First Defined
ε	Target suboptimality (accuracy)	Definition 1 (Preliminaries)
δ	Failure probability (confidence $1-\delta$)	Definition 1 (Preliminaries)
$N(\varepsilon, \delta)$	Episode budget for (ε, δ)-PAC	Definition 1 (Preliminaries)
K	Total episodes (regret horizon)	Preliminaries
$\tilde{\mathcal{O}}(\cdot), \tilde{\Theta}(\cdot), \tilde{\Omega}(\cdot)$	Suppress polylog factors	Taxonomy section
$\text{poly}(x)$	Polynomial in x (unspecified degree)	Throughout

Structural Complexity Parameters

Parameter	Meaning	Section
d	Feature dimension (linear MDPs)	Function Approximation
r	Low-rank factorization rank	Rich Observations
m	Latent state size (Block MDPs)	Rich Observations
B	Bellman rank	Structural Complexity
W	Witness rank	Structural Complexity
d_{BE}	Bellman-Eluder dimension	Structural Complexity
$d_{\mathrm{eff}}(\lambda)$	Effective dimension (RKHS/NTK)	Function Approximation
$\Phi_{\mathrm{tab}}(S, A, H)$	Tabular factor $\equiv SAH^3$	Definition 8 (Preliminaries)

Coverage and Occupancy

Symbol	Meaning	First Defined
$d_h^\pi(s, a)$	State-action occupancy of π at stage h	Definition 6 (Preliminaries)
$\mu_h(s, a)$	Behavior/data occupancy at stage h	Definition 6 (Preliminaries)
$C_\star$	Concentrability coefficient	Definition 6 (Preliminaries)

Function Classes and Operators

Symbol	Meaning	First Defined
$\mathcal{F}$	Value or Q-function class	Definition 3 (Preliminaries)
$\mathcal{M}$	Model class (dynamics hypotheses)	Structural Complexity
$\mathcal{G}$	Discriminator class (witness tests)	Structural Complexity
$\mathcal{T}_h$	Bellman optimality operator	Definition 5 (Preliminaries)
$\mathcal{N}(\epsilon, \mathcal{F}, \|\cdot\|)$	Covering number of $\mathcal{F}$	Definition 4 (Preliminaries)

Access Modes

- **Online:** Sequential interaction; agent chooses actions, observes transitions.
- **Generative model:** i.i.d. sampling from $P_h(\cdot \mid s, a)$ on query (s, a, h).
- **Offline:** Fixed dataset $\mathcal{D}$; no further interaction.

Abbreviations

Abbrev.	Meaning
RL	Reinforcement Learning
MDP	Markov Decision Process
POMDP	Partially Observable MDP
CDP	Contextual Decision Process
PAC	Probably Approximately Correct
RFE	Reward-Free Exploration
OPE	Off-Policy Evaluation
BPI	Best-Policy Identification
BE	Bellman-Eluder (dimension)
NTK	Neural Tangent Kernel
RKHS	Reproducing Kernel Hilbert Space
LSVI	Least-Squares Value Iteration
UCB	Upper Confidence Bound
PEVI	Pessimistic Episodic Value Iteration
PQL	Pessimistic Q-Learning
CSO	Coverage-Structure-Objective framework