Frozen Policy Iteration for MDPs with Stochastic Transitions under Linear Q^π Realizability

David Xia* Ruizhong Qiu† Hanghang Tong‡
University of Illinois Urbana-Champaign

April 2026

Abstract

We study sample and computation efficient online reinforcement learning under the linear Q^π realizability assumption in Markov Decision Processes (MDPs) with stochastic transitions. Prior algorithms in this setting either require computationally intractable optimization problems, rely on cost-sensitive oracles, or require local access to a simulator. Recent work introduced the Frozen Policy Iteration (FPI) algorithm, the first computationally efficient online procedure under linear Q^π realizability, but it is restricted to deterministic transitions and leaves the general, stochastic case as an open problem. FPI maintains, for each horizon stage, a dataset of state-action pairs whose Q-value estimates remain on-policy as the policy evolves. The key invariant is that a data pair is added to its dataset only after every state-action pair downstream of it is already *frozen*, locking in its recorded Q-value estimate. Under deterministic transitions, the trajectory descending from any state-action pair is unique, so this precondition is well-defined and easy to verify. Under stochastic transitions a single state-action pair branches into many possible downstream trajectories, and the precondition becomes uncheckable. We propose *Stochastic FPI*, which freezes policies on a *per-horizon* rather than per-state basis. Starting from the final horizon stage and working backwards, an entire horizon is frozen all at once only after every state-action pair at that horizon is well-explored. To collect data at horizons not yet frozen, we use a fixed reference policy π_{ref} supplied to the learner, paired with an early-stopping patience counter. Under a single-policy concentrability assumption with respect to π_{ref}, we prove a PAC bound of $\widetilde{O}(d^2H^7C^*/\varepsilon^3)$, where d is the feature dimension, H is the horizon length, C^* is the concentrability coefficient, and ε is the desired suboptimality.

1 Introduction

Reinforcement learning (RL) with function approximation has become a central topic in modern RL theory, with a primary goal of identifying structural assumptions on the function class that permit both statistically and computationally efficient learning. In the linear function approximation regime, two of the most studied assumptions are *linear Bellman completeness* [Golowich and Moitra, 2024, Wu et al., 2024, Mhammedi et al., 2026] and *linear Q^π realizability* [Du et al., 2019, Lattimore et al., 2020, Yin et al., 2022, Weisz et al., 2022, 2023]. The latter, which posits that every policy's Q-function is linear in a fixed feature map, has the desirable monotonicity property: enriching the feature class never breaks realizability. By contrast, the Bellman completeness assumption may

*<davidx3@illinois.edu>
†<rq5@illinois.edu>
‡<htong@illinois.edu>

break by adding features, which is a serious limitation for modern function approximators where features are learned and not hand-designed.

Until recently, all known sample-efficient algorithms either solved computationally intractable optimization problems [Weisz et al., 2023], relied on cost-sensitive classification oracles [Mhammedi, 2024], or required access to a simulator [Du et al., 2019, Lattimore et al., 2020, Yin et al., 2022, Weisz et al., 2022]. The recent Frozen Policy Iteration (FPI) algorithm of Ke et al. [2026] closed this gap for the special case of deterministic transitions, but left the extension to stochastic transitions as an open problem. In this paper we resolve this question affirmatively in the PAC setting, under an additional single-policy concentrability condition.

To set the stage for our contribution, we first describe FPI in some detail. FPI is a *policy-iteration* scheme. In every episode, the algorithm executes a current policy, observes a trajectory, and uses the trajectory to refine its Q-estimates and select a new policy for the next episode. The success of this scheme rests on a delicate invariant. Because the regression target at each visited state-action pair is the realized future return, and that return is governed by whatever policy was used when the trajectory was collected, the algorithm must somehow ensure the dataset remains *on-policy* as the policy evolves. FPI achieves this by *freezing* the policy at well-explored states. Data is added to a dataset only when every state-action pair downstream of it, along all possible trajectories it generates, is already *frozen* (locked in from an earlier episode and no longer subject to policy updates). Under deterministic transitions, this precondition is well-defined and easy to verify, because the trajectory descending from any state-action pair is unique. There is exactly one path of downstream states that must already be frozen before the data may be added. The Q-value estimates recorded at the newly added pair therefore reflects the now-fixed downstream behavior, and the dataset stays on-policy throughout learning.

Under stochastic transitions, the trajectory descending from a given state-action pair is no longer unique. A single execution from that pair samples just one of many possible downstream futures, and freezing the policy along the observed path leaves the policy free to change at the many alternative successor states the algorithm has not seen. The next time the algorithm visits the same pair, the transition kernel may deliver it to one of these unfrozen successors, and the future reward collected is no longer governed by the same policy that generated the regression target. The data is no longer on-policy, and the regression is no longer estimating any single Q-function.

The branching is not benign. At each subsequent horizon, the next state may be any of the potentially many states reachable in one transition, so the number of possible downstream trajectories from a single state-action pair grows multiplicatively with the horizon.

The structural fix is to freeze *horizon stages* rather than individual states. Starting from the final horizon of the MDP and working backwards, our algorithm freezes an entire horizon only after every state-action pair at that horizon has been well-explored. This sidesteps the branching problem entirely. By the time the algorithm begins collecting data at a particular horizon, every future horizon has already been frozen *as a whole*, so the policy at every conceivable successor state along any sampled trajectory is already fixed. The future return recorded at any added pair is therefore the same regardless of how the algorithm behaves elsewhere.

This raises a new difficulty. While the algorithm focuses on collecting data at the deepest unfrozen horizon, the earlier horizons must be traversed by some policy that does not depend on data the algorithm has not yet collected. We thus require that the learner is supplied with a single fixed *reference policy* π_{ref} for this purpose.

The state distribution at the deepest unfrozen horizon is now determined by π_{ref} on the early horizons. We have no direct control over this distribution, and it is possible that, even after many episodes, the trajectories rarely land at any informative state at the current horizon. Without a stopping rule, the algorithm could run indefinitely without progress. We therefore introduce

a *patience counter*. If a fixed number of consecutive episodes pass without adding new data at the current horizon, we declare that horizon frozen and move on. To convert this heuristic into a provable guarantee we appeal to single-policy concentrability with respect to π_{ref}, a standard offline-RL condition [Zhan et al., 2022, Zhu and Zhang, 2023, Tkachuk et al., 2024, Jiang and Xie, 2025] asserting that the optimal policy's state-action distribution at every horizon is bounded by a constant factor C^* above π_{ref}'s. This implies that any state-action pair the algorithm rarely visits under π_{ref} is also rare under π^*, so its contribution to the suboptimality is bounded.

Under linear Q^π realizability and single-policy concentrability with respect to π_{ref}, our *Stochastic FPI* algorithm finds an ε-optimal policy after at most $\widetilde{O}(d^2H^7C^*/\varepsilon^3)$ episodes, with running time polynomial in d, H, $|\mathcal{A}|$, T, $1/\varepsilon$.

2 Related Work

Linear Q^π realizability. The linear Q^π realizability setting was studied early by Du et al. [2019] and Lattimore et al. [2020], who gave polynomial-sample algorithms requiring access to a generative model. Yin et al. [2022] and Weisz et al. [2022] subsequently relaxed this requirement to local simulator access. Weisz et al. [2023] gave the first online algorithm with polynomial sample complexity, but their method employs a global optimism approach over large version spaces and is computationally intractable. Their analysis additionally requires the misspecification level κ to be at most $\widetilde{O}(\varepsilon^2/(d^6H^8))$. Mhammedi [2024] obtained an oracle-efficient algorithm using a cost-sensitive classification oracle, which can be NP-hard to implement in the worst case. Most recently, Ke et al. [2026] introduced Frozen Policy Iteration, the first computationally efficient online algorithm under linear Q^π realizability, restricted to deterministic transitions. Our work extends FPI to stochastic transitions at the cost of an additional concentrability assumption.

Misspecification. Du et al. [2019] prove that under linear Q^π realizability with misspecification $\kappa = \Omega(\sqrt{H/d})$, no algorithm can find a near-optimal policy with polynomially many samples, even given simulator access and knowing transitions. This places a fundamental ceiling on tolerable misspecification. The polynomial-sample upper bound of Weisz et al. [2023] requires $\kappa = \widetilde{O}(\varepsilon^2/(d^6H^8))$, leaving a substantial gap. Our algorithm tolerates $\kappa = \widetilde{O}(\varepsilon/(\sqrt{d}\,H))$, enough to absorb any polynomial dependence of κ on $d, H, 1/\varepsilon$ while remaining computationally efficient.

Concentrability and offline RL. Single-policy concentrability is a standard coverage condition in offline RL, where the constant C^* captures how well an offline data distribution covers the optimal policy's state-action visitation [Zhan et al., 2022, Zhu and Zhang, 2023, Tkachuk et al., 2024]. Recent work has begun importing this offline-style coverage assumption into the online setting. Xie et al. [2021] study *policy fine-tuning*, where the learner is given a reference policy μ that is concentrable against π^* and shows it can match the offline sample-complexity lower bound, while a hybrid algorithm can improve upon both purely offline and purely online RL. Xie et al. [2022] establish that the mere *existence* of a concentrable distribution, a property they term *coverability*, enables sample-efficient online exploration even when the distribution itself is unknown. Wagenmaker and Pacchiano [2023] characterize the optimal trade-off between offline data and online interaction in linear MDPs. Our use is closest in spirit to that of Xie et al. [2021]. The reference policy π_{ref} serves as a behavioral prior whose state-action visitation is required to dominate π^*'s up to a factor C^*. To our knowledge, this is the first use of single-policy concentrability with respect to a learner-supplied reference policy as a tool for online exploration in the linear Q^π realizability setting.

3 Preliminaries

3.1 Markov Decision Processes

A finite-horizon MDP is a tuple $(\mathcal{S}, \mathcal{A}, H, P, R, \mu)$, where $H \in \mathbb{N}^+$ is the horizon, $\mathcal{S} = \mathcal{S}_1 \cup \cdots \cup \mathcal{S}_H$ is the (disjoint) state space partitioned by stage, $\mathcal{A}$ is the action space, $P : \mathcal{S}_h \times \mathcal{A} \to \Delta(\mathcal{S}_{h+1})$ is the stochastic transition kernel, $R : \mathcal{S} \times \mathcal{A} \to \Delta([0,1])$ is the reward distribution with mean $r(s,a)$, and $\mu \in \Delta(\mathcal{S}_1)$ is the initial-state distribution. For a policy $\pi : \mathcal{S} \to \mathcal{A}$, the state-action and state value functions are

$$Q^\pi(s,a) = \mathbb{E}\left[\sum_{i=h}^{H} r_i \,\middle|\, s_h = s,\, a_h = a,\, \pi\right], \qquad V^\pi(s) = Q^\pi(s, \pi(s)),$$

for $s \in \mathcal{S}_h$. Let π^* denote an optimal policy. The *reference policy* π_{ref} is a fixed policy supplied to the learner. We assume the learner can compute or sample $\pi_{\mathrm{ref}}(s)$ for any $s \in \mathcal{S}$. For a positive-definite $V \in \mathbb{R}^{d\times d}$, the elliptical norm is $\|x\|_V = \sqrt{x^\top V x}$. We use $\widetilde{O}(\cdot)$ to suppress logarithmic factors.

3.2 Datasets and Timestamps

Throughout the paper, the algorithm maintains a per-horizon dataset

$$\mathcal{D}_h \;=\; \big\{(s_{h,i},\, a_{h,i},\, q_{h,i})\big\}_{i=1}^{D_h}, \qquad D_h := |\mathcal{D}_h|,$$

ordered by the episode in which each entry was appended. For each episode $t \geq 1$, we write $\mathcal{D}_{t,h}$ for the snapshot of $\mathcal{D}_h$ *at the start of episode* t (i.e., before any update in episode t has occurred), and we let

$$D_{t,h} := |\mathcal{D}_{t,h}|$$

denote its size, with $\mathcal{D}_{1,h} = \emptyset$ and $D_{1,h} = 0$. Because the datasets only grow during the algorithm, $\mathcal{D}_{t,h} \subseteq \mathcal{D}_{t+1,h}$ for every $t \geq 1$.

For each $h \in [H]$ and each $i \geq 1$, we define the *append timestamp*

$$t_{h,i} \;:=\; \min\big\{t \geq 1 :\, (s_{h,i}, a_{h,i}, q_{h,i}) \in \mathcal{D}_{t+1,h}\big\},$$

i.e., the episode in which the i-th entry of $\mathcal{D}_h$ was appended. We write $\phi_{h,i} := \phi(s_{h,i}, a_{h,i})$ for the feature of the i-th entry, and define the regularized empirical covariance restricted to the first k entries of $\mathcal{D}_h$ as

$$\Sigma_{h,k} \;:=\; \lambda I_d + \sum_{i=1}^{k} \phi_{h,i}\, \phi_{h,i}^\top, \qquad k \in \{0, 1, \ldots, D_h\},$$

so that $\Sigma_{h,0} = \lambda I_d$. We write $\Sigma_{t,h} := \Sigma_{h,D_{t,h}}$ for the covariance at the start of episode t.

3.3 Assumptions

Assumption 1 (Linear Q^π Realizability)**.** The MDP $(\mathcal{S}, \mathcal{A}, H, P, R, \mu)$ is κ-approximate Q^π-realizable with feature map $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$. That is, for any policy π, there exist $\theta_1^\pi, \ldots, \theta_H^\pi \in \mathbb{R}^d$ such that for any $h \in [H]$, $s \in \mathcal{S}_h$, $a \in \mathcal{A}$,

$$\big|Q^\pi(s,a) - \langle \phi(s,a), \theta_h^\pi \rangle\big| \leq \kappa.$$

Assumption 2 (Boundedness)**.** For all $h \in [H]$, $\|\phi(s,a)\|_2 \le 1$ for $s \in \mathcal{S}_h$, $a \in \mathcal{A}$, and $\|\theta_h^\pi\|_2 \le \sqrt{d}\,H$ for any policy π.

Assumption 3 (Single-policy concentrability with respect to π_{ref})**.** There exists a finite constant $C^* < \infty$ such that

$$C^* := \sup_{h\in[H],\,(s,a)\in\mathcal{S}_h\times\mathcal{A}} \frac{\mathbb{P}^{\pi^*}[s_h = s,\, a_h = a]}{\mathbb{P}^{\pi_{\mathrm{ref}}}[s_h = s,\, a_h = a]} < \infty,$$

with the convention $0/0 = 0$.

It is assumed that the feature map $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ in Assumption 1 is supplied to the learner and can be computed in poly(d) time. Assumption 3 is a standard condition in offline RL, where it captures the quality of an offline distribution relative to the optimal policy. Here we use it in an *online* setting. The data the algorithm collects under π_{ref} at unfrozen horizons plays a role analogous to an offline dataset for those horizons. The assumption implies that any state-action pair that is rare under π_{ref} is also rare under π^* (up to the factor C^*), which we exploit to bound the suboptimality from states the algorithm rarely visits. Choosing $\pi_{\mathrm{ref}} = \pi_{\mathrm{rand}}$ recovers a worst-case version of the assumption that always holds with $C^* \le |\mathcal{A}|^H$, but tighter coefficients are possible when the learner has access to a more informative reference policy (e.g., a known behavior policy or a domain expert).

4 Algorithm: Stochastic FPI

We now describe Stochastic FPI. The algorithm maintains the per-horizon datasets $\mathcal{D}_h$, snapshots $\mathcal{D}_{t,h}$, sizes $D_{t,h}$, and covariances $\Sigma_{t,h}$ of Section 3.2, where each entry $(s_{h,i}, a_{h,i}, q_{h,i}) \in \mathcal{D}_h$ has $q_{h,i}$ being the Q-value estimate, the cumulative reward from step h to H, collected from $(s_{h,i}, a_{h,i})$ under the policy $\pi_{t_{h,i}}$ in force when the entry was appended (cf. Section 3.2). The algorithm tracks a single state variable $h_u \in [H]$, the index of the *last non-frozen horizon*, initialized to $h_u = H$. Horizons $h > h_u$ are considered frozen. The algorithm is governed by an ellipsoid-tolerance parameter $\bar{\varepsilon} \in (0,1)$ and a patience tolerance $\mathcal{N} \in \mathbb{N}^+$, both fixed in advance.

Three policy phases. Within each episode t, the policy at step h depends on the relationship between h and h_u:

1. For $h > h_u$ (a frozen horizon): act greedily with respect to the least-squares Q-estimate built from the fixed dataset $\mathcal{D}_{t,h}$. This is the exploitation phase.
2. For $h = h_u$ (the active horizon): take the action with maximum information gain, $\pi_t(s) = \arg\max_{a\in\mathcal{A}} \|\phi(s,a)\|_{\Sigma_{t,h_u}^{-1}}$. This is the exploration phase.
3. For $h < h_u$ (an unexplored horizon): follow the reference policy, $\pi_t(s) = \pi_{\mathrm{ref}}(s)$. This is the data-routing phase, which serves only to deliver some state at horizon h_u.

Dataset updates. At the end of each episode t, if the action at h_u was sufficiently informative, namely

$$\big\|\phi\big(s_{h_u}^{(t)},\, a_{h_u}^{(t)}\big)\big\|_{\Sigma_{t,h_u}^{-1}} \;>\; \bar{\varepsilon},$$

we append $(s_{h_u}^{(t)}, a_{h_u}^{(t)}, \widehat{q}_{h_u}^{(t)})$ to $\mathcal{D}_{h_u}$, where $\widehat{q}_{h_u}^{(t)} := \sum_{h=h_u}^{H} r_h^{(t)}$ is the empirical Q-value estimate from step h_u onward in the trajectory of episode t. Otherwise, we increment a patience counter N, initialized to 0 at the start of each phase $h_u = h$.

Freezing rule. We declare h_u frozen and decrement it ($h_u \leftarrow h_u - 1$) whenever either of the following holds:

- Every $(s, a) \in \mathcal{S}_{h_u} \times \mathcal{A}$ is well-covered: $\|\phi(s,a)\|_{\Sigma_{t,h_u}^{-1}} < \bar{\varepsilon}$.
- The patience counter exceeds the threshold: $N > \mathcal{N}$.

The first criterion is the analog of the original FPI's freezing condition. The second criterion is new and is what makes the algorithm well-posed under stochastic transitions.

The full procedure is given as Algorithm 1.

Algorithm 1 Stochastic FPI – PAC

Require: ellipsoid tolerance $\bar{\varepsilon} \in (0,1)$, patience tolerance $\mathcal{N} \in \mathbb{N}^+$

For all $h \in [H]$, initialize $\mathcal{D}_h \leftarrow \emptyset$
Initialize last non-frozen horizon $h_u \leftarrow H$ and patience counter $N \leftarrow 0$
for episode $t = 1, \ldots, T$ **do**
  For each $h > h_u$ **and** $s \in \mathcal{S}_h$, define the greedy policy

$$\pi_t(s) = \arg\max_{a \in \mathcal{A}} Q_t(s,a) := \arg\max_{a \in \mathcal{A}} \Big\langle \phi(s,a),\, \Sigma_{t,h}^{-1} \sum_{i=1}^{D_{t,h}} \phi_{h,i}\, q_{h,i} \Big\rangle.$$

  For each $s \in \mathcal{S}_{h_u}$, define the exploratory policy $\pi_t(s) = \arg\max_{a \in \mathcal{A}} \|\phi(s,a)\|_{\Sigma_{t,h_u}^{-1}}$.
  For each $h < h_u$ **and** $s \in \mathcal{S}_h$, set $\pi_t(s) = \pi_{\text{ref}}(s)$
  Execute π_t and observe trajectory $(s_h^{(t)}, a_h^{(t)}, r_h^{(t)})_{h=1}^H$
  if $\|\phi(s_{h_u}^{(t)}, a_{h_u}^{(t)})\|_{\Sigma_{t,h_u}^{-1}} > \bar{\varepsilon}$ **then**
    Append $(s_{h_u}^{(t)}, a_{h_u}^{(t)}, \widehat{q}_{h_u}^{(t)})$ to $\mathcal{D}_{h_u}$, where $\widehat{q}_{h_u}^{(t)} = \sum_{h=h_u}^H r_h^{(t)}$
    $N \leftarrow 0$
  else
    $N \leftarrow N + 1$
  end if
  if $N > \mathcal{N}$ **or** $\|\phi(s,a)\|_{\Sigma_{t,h_u}^{-1}} < \bar{\varepsilon}$ for all $s \in \mathcal{S}_{h_u}, a \in \mathcal{A}$ **then**
    $h_u \leftarrow h_u - 1$; $N \leftarrow 0$
  end if
end for

Under the deterministic FPI of Ke et al. [2026], freezing was triggered per state (s, a). The rollout from (s, a) traces a unique downstream path that can be frozen point-by-point. Under stochastic transitions, there is no unique path. We combat this with the per-horizon construction. By the time we record any data at horizon h_u, every horizon $h > h_u$ has already been frozen by virtue of h_u's definition. Therefore, regardless of which $s_{h_u+1}^{(t)}, s_{h_u+2}^{(t)}, \ldots$ the stochastic transitions deliver, the policy at those states is fixed. This is the structural property formalized in Lemma 2.

Even when there exist informative states at h_u, the reference policy at horizons $h < h_u$ may not deliver us to them. Without a patience cap, the algorithm could continue forever, never adding data and never freezing. The patience cap $\mathcal{N}$ combined with the concentrability assumption converts "informative states are reached infrequently under π_{ref}" into "informative states are reached infrequently under π^*."

Following Ke et al. [2026] and assuming $\phi(s,a)$ can be computed in $\text{poly}(d)$ time, the algorithm has time complexity $\widetilde{O}(HT|\mathcal{A}|\,\text{poly}(d)/\bar{\varepsilon}^2)$ per episode for the exploitation step (computing

$\arg\max_a Q_t$) and $\widetilde{O}(HT|\mathcal{A}|\,\mathrm{poly}(d))$ overall.

5 Analysis

We now prove the PAC guarantee for Stochastic FPI. The proof structure initially follows the original FPI analysis: bound the dataset sizes (Lemma 1), establish the on-policy property (Lemma 2), prove a concentration bound for the empirical least-squares estimate (Lemma 6–8), bound the suboptimality of the policy on covered states (Lemma 9), and aggregate to obtain the main theorem. The new ingredients in our analysis are the patience lemma (Lemma 10), which controls the probability that the algorithm freezes prematurely, and the application of single-policy concentrability inside the performance-difference argument (Lemma 11).

5.1 Dataset Size Bound

Lemma 1. *For all $t \geq 1$ and $h \in [H]$,*

$$D_{t,h} \;\leq\; D \;:=\; \frac{2d}{\bar{\varepsilon}^2}\log\!\left(1+\frac{4\bar{\varepsilon}^{-4}}{\lambda^2}\right).$$

Proof. By the dataset-update rule of Algorithm 1 (line 9), an entry $(s_{h,i}, a_{h,i}, q_{h,i})$ is appended to $\mathcal{D}_h$ in episode $t_{h,i}$ only when

$$\left\|\phi(s_{h,i},a_{h,i})\right\|_{\Sigma_{h,i-1}^{-1}} \;>\; \bar{\varepsilon},$$

since $\Sigma_{t_{h,i},h} = \Sigma_{h,i-1}$ at the moment of appending. Letting $D_{t,h} =: m$, we have

$$\bar{\varepsilon}^{\,2}\cdot m \;\leq\; \sum_{i=1}^{m}\min\!\big(1,\,\|\phi_{h,i}\|^2_{\Sigma_{h,i-1}^{-1}}\big) \;\leq\; 2\log(\det\Sigma_{h,m}/\det\Sigma_{h,0})$$

by the elliptical potential lemma [Lattimore and Szepesvári, 2020]. The right-hand side is at most $d\log(1+m/(\lambda d))$, and bounding $\log(1+x)\leq\sqrt{x}$ for $x\geq 0$ and rearranging yields $m \leq 4\bar{\varepsilon}^{-4}d/\lambda$. Plugging this back into the elliptical-potential bound and rearranging gives $m \leq D$. □

5.2 The On-Policy Property of the Dataset

The next lemma is the structural counterpart of Lemma 2 in Ke et al. [2026], adapted to per-horizon freezing.

Lemma 2 (Freezing preserves on-policy returns)**.** *For any $h \in [H]$ and any $i \geq 1$, with $t_{h,i}$ the append timestamp of the i-th entry of $\mathcal{D}_h$ (Section 3.2),*

$$Q^{\pi_{t_{h,i}}}(s_{h,i},a_{h,i}) \;=\; Q^{\pi_{t'}}(s_{h,i},a_{h,i}) \qquad \textit{for all } t' \geq t_{h,i}.$$

Proof. Set $t := t_{h,i}$. At the moment of appending, the algorithm has $h_u = h$, so all horizons $h' > h$ are already frozen. Frozen horizons have their Q-functions and policies determined by datasets $\mathcal{D}_{h'}$ that no longer change. In any later episode $t' \geq t$, the policy at every state $s' \in \mathcal{S}_{h'}$ for $h' > h$ is therefore identical to that under π_t. Since $Q^\pi(s,a) = \mathbb{E}[r(s,a) + V^\pi(s') \mid s,a]$ depends only on the policy at horizons $> h$, we conclude $Q^{\pi_t}(s_{h,i},a_{h,i}) = Q^{\pi_{t'}}(s_{h,i},a_{h,i})$ as claimed. □

5.3 Martingale and Subgaussian Properties of the Noise

Lemma 3. *Let $\xi_t := \widehat{q}_t - Q^{\pi_t}(s_{h_u}^{(t)}, a_{h_u}^{(t)})$, where h_u is the last non-frozen horizon at episode t. Let $\{\mathcal{F}_t\}_{t\geq 0}$ be the filtration with $\mathcal{F}_t$ generated by $\left(s_h^{(i)}, a_h^{(i)}, r_h^{(i)}\right)_{1\leq i\leq t, h\in[H]}$. Then $\{\xi_t\}_{t\geq 1}$ is a martingale difference sequence with respect to $\{\mathcal{F}_t\}$.*

Proof. Measurability. $\widehat{q}_t = \sum_{h=h_u}^{H} r_h^{(t)}$ is a function of within-episode rewards and is therefore $\mathcal{F}_t$-measurable. The policy π_t is determined by the datasets at the start of episode t and is $\mathcal{F}_{t-1}$-measurable; the horizon h_u is similarly $\mathcal{F}_{t-1}$-measurable. The state $s_{h_u}^{(t)}$ and action $a_{h_u}^{(t)}$ are $\mathcal{F}_t$-measurable. Since $Q^{\pi_t}(s,a)$ is a deterministic function of π_t, s, and a, the term $Q^{\pi_t}(s_{h_u}^{(t)}, a_{h_u}^{(t)})$ is $\mathcal{F}_t$-measurable, so ξ_t is $\mathcal{F}_t$-measurable.

Mean-zero. By the tower property,

$$\mathbb{E}[\xi_t \mid \mathcal{F}_{t-1}] = \mathbb{E}\Big[\mathbb{E}[\xi_t \mid \mathcal{F}_{t-1}, s_{h_u}^{(t)}, a_{h_u}^{(t)}] \,\Big|\, \mathcal{F}_{t-1}\Big].$$

Given $\mathcal{F}_{t-1}$, $s_{h_u}^{(t)}$, and $a_{h_u}^{(t)}$, the value $Q^{\pi_t}(s_{h_u}^{(t)}, a_{h_u}^{(t)})$ is a deterministic constant, while the remaining randomness in $\widehat{q}_t$ comes from the stochastic transitions and rewards from step h_u onward. By definition of Q^{π_t}, $\mathbb{E}[\widehat{q}_t \mid \mathcal{F}_{t-1}, s_{h_u}^{(t)}, a_{h_u}^{(t)}] = Q^{\pi_t}(s_{h_u}^{(t)}, a_{h_u}^{(t)})$, so the inner conditional expectation is zero. The tower property gives $\mathbb{E}[\xi_t \mid \mathcal{F}_{t-1}] = 0$. □

Lemma 4. *For each $t \geq 1$, $\xi_t \mid \mathcal{F}_{t-1}$ is H-subgaussian.*

Proof. Both $\widehat{q}_t \in [0, H]$ and $Q^{\pi_t}(s_{h_u}^{(t)}, a_{h_u}^{(t)}) \in [0, H]$, so $\xi_t \in [-H, H]$. By Hoeffding's lemma, any random variable bounded in $[a, b]$ is $(b-a)/2$-subgaussian, so $\xi_t \mid \mathcal{F}_{t-1}$ is H-subgaussian. □

For each $h \in [H]$ and $i \geq 1$, we extract from this sequence the noise term associated with the i-th entry of $\mathcal{D}_h$:

$$\xi_{h,i} \;:=\; \xi_{t_{h,i}} \;=\; q_{h,i} - Q^{\pi_{t_{h,i}}}(s_{h,i}, a_{h,i}),$$

where the second equality uses that, by definition of the append timestamp $t_{h,i}$, the active horizon during episode $t_{h,i}$ is $h_u = h$ and the recorded Q-value estimate is $q_{h,i} = \widehat{q}_{h_u}^{(t_{h,i})}$.

Remark 1. In the deterministic-transition analysis of Ke et al. [2026], ξ_t is a sum of H independent 1-subgaussians, giving $\sqrt{H}$-a subgaussian result. Under stochastic transitions, individual stage rewards along the trajectory are no longer independent of one another conditional on $\mathcal{F}_{t-1}$ (because $s_h^{(t)}$ is correlated with $s_{h-1}^{(t)}$ through P), so this variance reduction is no longer available. We therefore lose a factor of $\sqrt{H}$ in the noise scale, which propagates to the regret bound.

5.4 Concentration Bound

Lemma 5 (Self-normalized concentration; Abbasi-Yadkori et al., 2011)**.** *Let $\{\eta_t\}_{t\geq 1}$ be an MDS with respect to filtration $\{\mathcal{F}_t\}_{t\geq 0}$ with $\eta_t \mid \mathcal{F}_{t-1}$ being σ-subgaussian. Let $\{\phi_t\}_{t\geq 1}$ be an $\mathbb{R}^d$-valued process with ϕ_t being $\mathcal{F}_{t-1}$-measurable. Let Λ_0 be a positive-definite $d \times d$ matrix and $\Lambda_t = \Lambda_0 + \sum_{i=1}^{t} \phi_i \phi_i^\top$. Then for any $\delta' > 0$, with probability at least $1 - \delta'$, simultaneously for all $t \geq 1$,*

$$\Big\| \sum_{i=1}^{t} \phi_i \eta_i \Big\|_{\Lambda_t^{-1}}^2 \leq 2\sigma^2 \log\left(\frac{\det(\Lambda_t)^{1/2} \det(\Lambda_0)^{-1/2}}{\delta'} \right).$$

Lemma 6. *Define the event*

$$\mathcal{E}_{\text{high}} = \left\{ \forall h \in [H],\, \forall k \in \{1, \dots, D\} : \left\| \sum_{i=1}^{k} \phi_{h,i}\, \xi_{h,i} \right\|^2_{\Sigma_{h,k}^{-1}} \le 2H^2 \big(\tfrac{d}{2} \log\big(1 + \tfrac{k}{d\lambda}\big) + \log \tfrac{2H}{\delta} \big) \right\}.$$

Then $\mathbb{P}[\mathcal{E}_{\text{high}}] \ge 1 - \delta/2$.

Proof. Fix $h \in [H]$. By Lemmas 3 and 4, $\{\xi_{h,i}\}_{i \ge 1}$ is an H-subgaussian MDS, and $\{\phi_{h,i}\}_{i \ge 1}$ is an $\mathbb{R}^d$-valued sequence adapted to that filtration. Lemma 5 with $\sigma = H$, $\Lambda_0 = \lambda I_d$, $\delta' = \delta/(2H)$, and the standard determinant bound $\det(\Sigma_{h,k}) \le \big((d\lambda + k)/d\big)^d$ (using $\|\phi\|_2 \le 1$ from Assumption 2) yields the per-h bound. A union bound over $h \in [H]$ completes the proof. □

Lemma 7 (Zanette et al., 2020)**.** *For any sequence* $\{\delta_i\}_{i=1}^n \subseteq \mathbb{R}$ *with* $|\delta_i| \le \kappa$ *and* $\{\phi_i\}_{i=1}^n \subseteq \mathbb{R}^d$*, letting* $\Lambda = \lambda I_d + \sum_{i=1}^n \phi_i \phi_i^\top$*, we have* $\big\| \sum_{i=1}^n \phi_i \delta_i \big\|^2_{\Lambda^{-1}} \le n\kappa^2$.

5.5 Least-Squares Error Bound

Lemma 8. *Under* $\mathcal{E}_{\text{high}}$*, for any* $t \ge 1$*,* $h \in [H]$*,* $s \in \mathcal{S}_h$*,* $a \in \mathcal{A}$*,*

$$\big| Q_t(s,a) - Q^{\pi_t}(s,a) \big| \le \alpha \, \|\phi(s,a)\|_{\Sigma_{t,h}^{-1}} + \kappa,$$

where

$$\alpha = \sqrt{2H^2 \big(\tfrac{d}{2} \log\big(1 + \tfrac{D}{\lambda d}\big) + \log \tfrac{2H}{\delta} \big)} + \sqrt{D}\,\kappa + \sqrt{\lambda d}\, H,$$

and $\Sigma_{t,h} = \lambda I_d + \sum_{i=1}^{D_{t,h}} \phi_{h,i} \phi_{h,i}^\top$.

Proof. Let $n := D_{t,h}$ and $\delta_i := Q^{\pi_t}(s_{h,i}, a_{h,i}) - \langle \phi_{h,i}, \theta_h^{\pi_t} \rangle$, so $|\delta_i| \le \kappa$ by Assumption 1. By the triangle inequality,

$$|Q_t(s,a) - Q^{\pi_t}(s,a)| \le |Q_t(s,a) - \langle \phi(s,a), \theta_h^{\pi_t} \rangle| + \kappa,$$

so it suffices to bound the first term. Recall $Q_t(s,a) = \big\langle \phi(s,a),\, \Sigma_{t,h}^{-1} \sum_{i=1}^n \phi_{h,i}\, q_{h,i} \big\rangle$. Since $t_{h,i} \le t$ for entries already in $\mathcal{D}_{t,h}$, by Lemma 2, $Q^{\pi_{t_{h,i}}}(s_{h,i}, a_{h,i}) = Q^{\pi_t}(s_{h,i}, a_{h,i})$ for every $i \in \{1, \dots, n\}$. Thus, we have that

$$q_{h,i} \;=\; Q^{\pi_{t_{h,i}}}(s_{h,i}, a_{h,i}) + \xi_{h,i} \;=\; Q^{\pi_t}(s_{h,i}, a_{h,i}) + \xi_{h,i} \;=\; \langle \phi_{h,i}, \theta_h^{\pi_t} \rangle + \delta_i + \xi_{h,i}.$$

Substituting,

$$\begin{aligned} Q_t(s,a) - \langle \phi(s,a), \theta_h^{\pi_t} \rangle &= \Big\langle \phi(s,a),\, \Sigma_{t,h}^{-1} \sum_{i=1}^n \phi_{h,i} \big(\langle \phi_{h,i}, \theta_h^{\pi_t} \rangle + \delta_i + \xi_{h,i} \big) - \theta_h^{\pi_t} \Big\rangle \\ &= \Big\langle \phi(s,a),\, \Sigma_{t,h}^{-1} \sum_{i=1}^n \phi_{h,i} (\delta_i + \xi_{h,i}) - \lambda \Sigma_{t,h}^{-1} \theta_h^{\pi_t} \Big\rangle, \end{aligned}$$

using $\sum_i \phi_{h,i} \phi_{h,i}^\top = \Sigma_{t,h} - \lambda I_d$. Applying the generalized Cauchy–Schwarz inequality $|\langle u, v \rangle| \le \|u\|_M \|v\|_{M^{-1}}$ with $M = \Sigma_{t,h}$,

$$|Q_t(s,a) - \langle \phi(s,a), \theta_h^{\pi_t} \rangle| \le \|\phi(s,a)\|_{\Sigma_{t,h}^{-1}} \Big(\Big\| \sum_{i=1}^n \phi_{h,i} \xi_{h,i} \Big\|_{\Sigma_{t,h}^{-1}} + \Big\| \sum_{i=1}^n \phi_{h,i} \delta_i \Big\|_{\Sigma_{t,h}^{-1}} + \lambda \|\theta_h^{\pi_t}\|_{\Sigma_{t,h}^{-1}} \Big).$$

Under $\mathcal{E}_{\text{high}}$, the first term is at most $\sqrt{2H^2 (\frac{d}{2} \log(1 + \frac{n}{\lambda d}) + \log \frac{2H}{\delta})}$ by Lemma 6; the second is at most $\sqrt{n}\,\kappa$ by Lemma 7; and the third is at most $\sqrt{\lambda}\, \|\theta_h^{\pi_t}\|_2 \le \sqrt{\lambda d}\, H$ by $\Sigma_{t,h}^{-1} \preceq \lambda^{-1} I_d$ and Assumption 2. Since $n \le D$ by Lemma 1, the sum is bounded by α, and the lemma follows. □

Lemma 9. *Under $\mathcal{E}_{\text{high}}$, for any $t \geq 1$, $h \in [H]$, $s \in \mathcal{S}_h$, $a \in \mathcal{A}$, if $\|\phi(s,a)\|_{\Sigma_{t,h}^{-1}} \leq \bar{\varepsilon}$, then $|Q_t(s,a) - Q^{\pi_t}(s,a)| \leq \alpha\bar{\varepsilon} + \kappa$.*

Proof. Direct from Lemma 8. □

5.6 The Patience Lemma

Lemma 10 (Patience)**.** *Fix horizon $h \in [H]$. During any phase in which $h_u = h$, all successor horizons $h' > h$ are frozen, so the dataset $\mathcal{D}_{t,h}$, the covariance $\Sigma_{t,h}$, and the covered set*

$$\text{Cover}(\bar{\varepsilon}) \;:=\; \left\{ s \in \mathcal{S}_h : \|\phi(s,a)\|_{\Sigma_{t,h}^{-1}} < \bar{\varepsilon} \text{ for all } a \in \mathcal{A} \right\}$$

are all fixed throughout any window of consecutive episodes during which no new data is appended to $\mathcal{D}_h$.

Suppose during such a phase $\mathcal{N}$ consecutive episodes pass without any data being appended to $\mathcal{D}_h$. Then with probability at least $1 - \delta/(2H)$,

$$\mathbb{P}^{\pi_{\text{ref}}}\big[s_h \notin \text{Cover}(\bar{\varepsilon})\big] \;\leq\; \frac{\log(2H/\delta)}{\mathcal{N}}.$$

By a union bound over $h \in [H]$, this holds simultaneously for every horizon with probability $\geq 1 - \delta/2$.

Proof. Fix the phase $h_u = h$ and a window of $\mathcal{N}$ consecutive episodes within it during which no data is appended. Throughout the window, $\Sigma_{t,h}$ does not change, so $\text{Cover}(\bar{\varepsilon})$ is a fixed set $\mathcal{C} \subseteq \mathcal{S}_h$. While $h_u = h$, the policy at every $h' < h$ is π_{ref}, while at $h' = h$ the policy chooses an action via $\arg\max_{a \in \mathcal{A}} \|\phi(s,a)\|_{\Sigma_{t,h}^{-1}}$. The marginal distribution of $s_h^{(t)}$ across these episodes is therefore exactly $\mathbb{P}^{\pi_{\text{ref}}}[s_h \in \cdot\,]$, since the choice of policy at h_u does not affect the state reached *at* h_u.

Let $p := \mathbb{P}^{\pi_{\text{ref}}}[s_h \notin \mathcal{C}]$. Across the $\mathcal{N}$ episodes, the $\mathcal{N}$ events $\{s_h^{(t)} \in \mathcal{C}\}$ are i.i.d. Bernoulli$(1-p)$ since the dataset is fixed and the underlying distribution is the same in each episode. The probability that all $\mathcal{N}$ episodes land in $\mathcal{C}$ is therefore $(1-p)^{\mathcal{N}}$. If $p \geq \log(2H/\delta)/\mathcal{N}$, this probability is at most $\exp(-\log(2H/\delta)) = \delta/(2H)$. Contrapositively, with probability at least $1 - \delta/(2H)$ we have $p \leq \log(2H/\delta)/\mathcal{N}$. A union bound over horizons completes the proof. □

5.7 Near-Optimality of the Frozen Policy

Lemma 11 (Near-optimality)**.** *Condition on $\mathcal{E}_{\text{high}}$ and the patience event of Lemma 10 (jointly holding with probability $\geq 1 - \delta$). For any episode t at which all H horizons are frozen, and for any $s \in \mathcal{S}_1$,*

$$V^{\pi^*}(s) - V^{\pi_t}(s) \;\leq\; 2H(\alpha\bar{\varepsilon} + \kappa) + \frac{H^2 C^* \log(2H/\delta)}{\mathcal{N}}.$$

Proof. By the performance difference lemma [Kakade and Langford, 2002],

$$V^{\pi^*}(s) - V^{\pi_t}(s) = \sum_{h=1}^{H} \mathbb{E}\big[\, Q^{\pi_t}(s_h^*, \pi^*(s_h^*)) - Q^{\pi_t}(s_h^*, \pi_t(s_h^*)) \,\big],$$

where the expectation is over the stochastic trajectory $s_1^*, s_2^*, \ldots$ of π^* starting at $s_1^* = s$. We bound each summand by partitioning on whether $s_h^* \in \text{Cover}(\bar{\varepsilon})$.

Case 1: $s_h^* \in \text{Cover}(\bar{\varepsilon})$. For every $a \in \mathcal{A}$, $\|\phi(s_h^*, a)\|_{\Sigma_{t,h}^{-1}} < \bar{\varepsilon}$, so by Lemma 9, $|Q_t(s_h^*, a) - Q^{\pi_t}(s_h^*, a)| \le \alpha\bar{\varepsilon} + \kappa$. Since $\pi_t(s_h^*) = \arg\max_a Q_t(s_h^*, a)$,

$$\begin{aligned} Q^{\pi_t}(s_h^*, \pi^*(s_h^*)) &\le Q_t(s_h^*, \pi^*(s_h^*)) + (\alpha\bar{\varepsilon} + \kappa) \\ &\le Q_t(s_h^*, \pi_t(s_h^*)) + (\alpha\bar{\varepsilon} + \kappa) \\ &\le Q^{\pi_t}(s_h^*, \pi_t(s_h^*)) + 2(\alpha\bar{\varepsilon} + \kappa). \end{aligned}$$

Therefore the summand is at most $2(\alpha\bar{\varepsilon} + \kappa)$.

Case 2: $s_h^* \notin \text{Cover}(\bar{\varepsilon})$. The summand is at most H since $Q^{\pi_t} \in [0, H]$. The probability of this case under π^* is bounded by Assumption 3 and Lemma 10:

$$\mathbb{P}^{\pi^*}[s_h^* \notin \text{Cover}(\bar{\varepsilon})] \;\le\; C^* \cdot \mathbb{P}^{\pi_{\text{ref}}}[s_h \notin \text{Cover}(\bar{\varepsilon})] \;\le\; \frac{C^* \log(2H/\delta)}{\mathcal{N}}.$$

Combining the two cases,

$$\mathbb{E}[Q^{\pi_t}(s_h^*, \pi^*(s_h^*)) - Q^{\pi_t}(s_h^*, \pi_t(s_h^*))] \le 2(\alpha\bar{\varepsilon} + \kappa) + \frac{HC^* \log(2H/\delta)}{\mathcal{N}}.$$

Summing over $h \in [H]$ yields the claim. □

5.8 Main Theorem

Theorem 1 (PAC bound for Stochastic FPI)**.** *There is an absolute constant B such that for any target accuracy $\varepsilon > B\sqrt{d}\, H\kappa$, by setting*

$$\bar{\varepsilon} \;=\; \frac{\varepsilon - 4H\kappa}{4H\alpha}, \qquad \mathcal{N} \;=\; \left\lceil \frac{2H^2 C^* \log(2H/\delta)}{\varepsilon} \right\rceil,$$

with probability at least $1 - \delta$, the number of episodes whose suboptimality gap exceeds ε is at most

$$\widetilde{O}\left(\frac{d^2 H^7 C^*}{\varepsilon^3}\right).$$

Proof. We work on the joint event of $\mathcal{E}_{\text{high}}$ and the patience event of Lemma 10 of probability at least $1 - \delta$.

Optimal episodes. Once all H horizons are frozen, Lemma 11 bounds the suboptimality by $2H(\alpha\bar{\varepsilon} + \kappa) + H^2 C^* \log(2H/\delta)/\mathcal{N}$. We balance the two terms to make each at most $\varepsilon/2$. Setting $2H(\alpha\bar{\varepsilon} + \kappa) \le \varepsilon/2$ yields $\bar{\varepsilon} \le (\varepsilon - 4H\kappa)/(4H\alpha)$, and setting $H^2 C^* \log(2H/\delta)/\mathcal{N} \le \varepsilon/2$ yields $\mathcal{N} \ge 2H^2 C^* \log(2H/\delta)/\varepsilon$.

Suboptimal episodes. A suboptimal episode (gap $> \varepsilon$) can occur only before all horizons are frozen, i.e. , during a phase $h_u \in [H]$. Each such phase ends in one of two ways. It either accumulates new data (*informative* episodes) or it accumulates patience (*uninformative* episodes that increment N).

Informative episodes. Each dataset has size at most $D = \widetilde{O}(d/\bar{\varepsilon}^2)$ by Lemma 1, so across all H horizons at most HD informative episodes occur.

Uninformative episodes. Within a single phase $h_u = h$, the patience counter is reset to 0 after each informative episode and capped at $\mathcal{N}$. Therefore, between any two consecutive informative episodes, there are at most $\mathcal{N}$ uninformative episodes. Across all H phases, we have at most $H \cdot D \cdot \mathcal{N}$ episodes up to constants.

Computing the bound. Set $\lambda = H^{-1}$. The dominant term in α is $\sqrt{2H^2 \cdot \frac{d}{2}\log(\cdot)} = \widetilde{O}(H\sqrt{d})$. With $\bar{\varepsilon} = \Theta(\varepsilon/(H\alpha)) = \Theta(\varepsilon/(H^2\sqrt{d}))$,

$$H \cdot D \;=\; \widetilde{O}\Big(\frac{d \cdot H}{\bar{\varepsilon}^{\,2}}\Big) \;=\; \widetilde{O}\Big(\frac{d^2H^5}{\varepsilon^{\,2}}\Big), \qquad \mathcal{N} \;=\; \widetilde{O}\Big(\frac{H^2C^*}{\varepsilon}\Big).$$

Therefore the total number of bad episodes is $H \cdot D \cdot \mathcal{N} = \widetilde{O}(d^2H^7C^*/\varepsilon^{\,3})$. □

6 Comparison and Discussion

Table 1: PAC bounds for online RL under linear Q^π realizability.

Algorithm	**PAC bound**	**Notes**
Weisz et al. [2023]	$\widetilde{O}(d^7H^{11}/\varepsilon^2)$	Computationally intractable
FPI [Ke et al., 2026]	$\widetilde{O}(d^2H^4/\varepsilon^2)$	Deterministic transitions
This work	$\widetilde{O}(d^2H^7C^*/\varepsilon^3)$	Concentrability assumption

The role of concentrability. Assumption 3 is the price we pay for handling stochastic transitions without a simulator. The intuition is that, without a simulator, the only way to reach a given state at horizon h is to traverse the actual transition kernel from the initial distribution. If some state is unreachable under any policy of bounded behavior, we cannot test or learn there. Single-policy concentrability with respect to π_{ref} rules out the worst case in which the optimal policy concentrates mass on regions that π_{ref} never visits.

Choosing the reference policy. The reference policy π_{ref} enters our analysis only through (i) the patience lemma (Lemma 10), which relates the rate of new data to the marginal $\mathbb{P}^{\pi_{\text{ref}}}[s_h \in \cdot\,]$, and (ii) the concentrability constant C^*. Two convenient properties make any fixed π_{ref} analytically clean: its state-visitation distribution at each horizon does not change as the algorithm collects data, and the resulting C^* has the standard offline-RL interpretation. The default choice $\pi_{\text{ref}} = \pi_{\text{rand}}$ requires no prior knowledge but yields a worst-case $C^* \le |\mathcal{A}|^H$. When the learner has access to a more informative reference policy – for example, a known behavior policy that produced offline data, an expert demonstrator, or a coarse heuristic for the task – the concentrability coefficient can be substantially smaller, directly reducing the sample complexity. An interesting future direction is whether π_{ref} can itself be *learned* online without circular dependence on the dataset the algorithm is trying to build.

Open problems. Several questions remain. First, can the concentrability assumption be removed entirely, perhaps by integrating an optimism-style mechanism with the per-horizon freezing? Second, is the $1/\varepsilon^{\,3}$ rate tight, or is the patience overhead an artifact of the analysis?

7 Conclusion

We have introduced Stochastic FPI, the first computationally efficient online RL algorithm under linear Q^π realizability for MDPs with stochastic transitions, and proven a PAC guarantee of

$\widetilde{O}(d^2H^7C^*/\varepsilon^3)$ under single-policy concentrability with respect to a learner-supplied reference policy π_{ref}. Our main conceptual contributions are (i) per-horizon freezing as a structural replacement for the per-state freezing of Ke et al. [2026], which keeps the dataset on-policy under stochastic dynamics, and (ii) a patience-driven early-stopping rule that ensures progress when the reference policy is slow to deliver informative states. Closing the gap between our PAC rate and that of the deterministic FPI, removing the concentrability assumption, and tightening the patience analysis are all exciting directions for future work.

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24, 2011.

Simon S Du, Sham M Kakade, Ruosong Wang, and Lin F Yang. Is a good representation sufficient for sample efficient reinforcement learning? *arXiv preprint arXiv:1910.03016*, 2019.

Noah Golowich and Ankur Moitra. Linear bellman completeness suffices for efficient online reinforcement learning with few actions. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 1939–1981. PMLR, 2024.

Nan Jiang and Tengyang Xie. Offline reinforcement learning in large state spaces: Algorithms and guarantees. 2024. *URL https://arxiv. org/abs/2510.04088*, 2025.

Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *Proceedings of the nineteenth international conference on machine learning*, pages 267–274, 2002.

Yijing Ke, Zihan Zhang, and Ruosong Wang. Frozen policy iteration: Computationally efficient rl under linear q^π realizability for deterministic dynamics. *arXiv preprint arXiv:2603.00716*, 2026.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Tor Lattimore, Csaba Szepesvari, and Gellert Weisz. Learning with good feature representations in bandits and in rl with a generative model. In *International conference on machine learning*, pages 5662–5670. PMLR, 2020.

Zakaria Mhammedi. Sample and oracle efficient reinforcement learning for mdps with linearly-realizable value functions. *arXiv preprint arXiv:2409.04840*, 2024.

Zakaria Mhammedi, Alexander Rakhlin, and Nneka Okolo. End-to-end efficient rl for linear bellman complete mdps with deterministic transitions. *arXiv preprint arXiv:2603.23461*, 2026.

Volodymyr Tkachuk, Gellért Weisz, and Csaba Szepesvári. Trajectory data suffices for statistically efficient learning in offline rl with linear q^π-realizability and concentrability. *Advances in Neural Information Processing Systems*, 37:83268–83313, 2024.

Andrew Wagenmaker and Aldo Pacchiano. Leveraging offline data in online reinforcement learning. In *International Conference on Machine Learning*, pages 35300–35338. PMLR, 2023.

Gellért Weisz, András György, Tadashi Kozuno, and Csaba Szepesvári. Confident approximate policy iteration for efficient local planning in q^π-realizable mdps. *Advances in Neural Information Processing Systems*, 35:25547–25559, 2022.

Gellért Weisz, András György, and Csaba Szepesvári. Online rl in linearly q^π-realizable mdps is as easy as in linear mdps if you learn what to ignore. *Advances in Neural Information Processing Systems*, 36:59172–59205, 2023.

Runzhe Wu, Ayush Sekhari, Akshay Krishnamurthy, and Wen Sun. Computationally efficient rl under linear bellman completeness for deterministic dynamics. *arXiv preprint arXiv:2406.11810*, 2024.

Tengyang Xie, Nan Jiang, Huan Wang, Caiming Xiong, and Yu Bai. Policy finetuning: Bridging sample-efficient offline and online reinforcement learning. *Advances in neural information processing systems*, 34:27395–27407, 2021.

Tengyang Xie, Dylan J Foster, Yu Bai, Nan Jiang, and Sham M Kakade. The role of coverage in online reinforcement learning. *arXiv preprint arXiv:2210.04157*, 2022.

Dong Yin, Botao Hao, Yasin Abbasi-Yadkori, Nevena Lazić, and Csaba Szepesvári. Efficient local planning with linear function approximation. In *International Conference on Algorithmic Learning Theory*, pages 1165–1192. PMLR, 2022.

Andrea Zanette, Alessandro Lazaric, Mykel Kochenderfer, and Emma Brunskill. Learning near optimal policies with low inherent bellman error. In *International Conference on Machine Learning*, pages 10978–10989. PMLR, 2020.

Wenhao Zhan, Baihe Huang, Audrey Huang, Nan Jiang, and Jason Lee. Offline reinforcement learning with realizability and single-policy concentrability. In *Conference on Learning Theory*, pages 2730–2775. PMLR, 2022.

Hanlin Zhu and Amy Zhang. Provably efficient offline goal-conditioned reinforcement learning with general function approximation and single-policy concentrability. *Advances in Neural Information Processing Systems*, 36:4177–4198, 2023.