Lecture 11: Strategic Exploration

This lecture studies the *exploration problem* in the *finite-horizon episodic* setting. The agent interacts with an unknown MDP over K episodes, adaptively choosing policies, and the goal is to minimize *regret*. The central principle is *optimism in the face of uncertainty* (OFU). We cover three algorithms: *UCBVI* for tabular MDPs (pointwise optimism via count-based bonuses), *Lin-UCBVI* for linear MDPs (ellipsoidal bonuses), and GOLF for general function classes (global optimism via confidence sets).

Setup. We work with a *finite-horizon episodic MDP* $M = (\mathcal{S}, \mathcal{A}, H, \{P_h^\star\}_{h=0}^{H-1}, \{r_h\}_{h=0}^{H-1}, s_0)$, where $\mathcal{S}$ is a finite state space with $|\mathcal{S}| = S$, $\mathcal{A}$ is a finite action space with $|\mathcal{A}| = A$, H is the horizon, $P_h^\star(\cdot \mid s, a)$ is the (stage-dependent) transition kernel, $r_h(s, a) \in [0, 1]$ is a deterministic reward, and s_0 is a fixed initial state. A policy is $\pi = \{\pi_h : \mathcal{S} \to \mathcal{A}\}_{h=0}^{H-1}$. Value functions are defined stage-by-stage:

$$V_h^\pi(s) := \mathbb{E}\left[\sum_{t=h}^{H-1} r_t(s_t, a_t) \,\middle|\, s_h = s,\ \pi,\ P^\star\right], \quad Q_h^\pi(s, a) := r_h(s, a) + \mathbb{E}_{s' \sim P_h^\star(\cdot|s,a)}[V_{h+1}^\pi(s')].$$

We write $V_h^\star = \max_\pi V_h^\pi$ and $Q_h^\star = \max_\pi Q_h^\pi$ for the optimal value functions, and set $V_H^\star \equiv 0$.

Finite-horizon Bellman optimality equations:

$$Q_h^\star(s, a) = r_h(s, a) + P_h^\star V_{h+1}^\star(s, a), \qquad V_h^\star(s) = \max_{a \in \mathcal{A}} Q_h^\star(s, a), \qquad V_H^\star \equiv 0,$$

where we use the shorthand $(P_h^\star f)(s, a) := \mathbb{E}_{s' \sim P_h^\star(\cdot|s,a)}[f(s')]$ for any function $f : \mathcal{S} \to \mathbb{R}$.

Since rewards lie in $[0, 1]$, we have $0 \le V_h^\pi(s) \le H - h$ for all (h, s, π). For any policy π, the *stage-h occupancy measure* is $d_h^\pi(s, a) := \Pr(s_h = s, a_h = a \mid s_0, \pi, P^\star)$.

The Online Exploration Problem

The Episodic Protocol and Regret

A learning algorithm interacts with the MDP over K episodes. In each episode $k = 0, 1, \ldots, K - 1$:

1. The algorithm selects a policy $\pi^k = \{\pi_h^k\}_{h=0}^{H-1}$ based on all data collected in episodes $0, \ldots, k - 1$.
2. The algorithm executes π^k for one episode: starting from $s_0^k = s_0$, it takes action

$a_h^k = \pi_h^k(s_h^k)$, observes reward $r_h(s_h^k, a_h^k)$, and transitions to $s_{h+1}^k \sim P_h^\star(\cdot \mid s_h^k, a_h^k)$, for $h = 0, \ldots, H-1$.

3. The algorithm records the trajectory $(s_0^k, a_0^k, s_1^k, a_1^k, \ldots, s_{H-1}^k, a_{H-1}^k, s_H^k)$.

The performance measure is the cumulative regret:

Definition 1 (Regret)**.** *The* (cumulative) regret *over K episodes is*

$$R_K \coloneqq \sum_{k=0}^{K-1} \left[V_0^\star(s_0) - V_0^{\pi^k}(s_0)\right].$$

Regret:

$$R_K = \sum_{k=0}^{K-1} \left[V_0^\star(s_0) - V_0^{\pi^k}(s_0)\right].$$

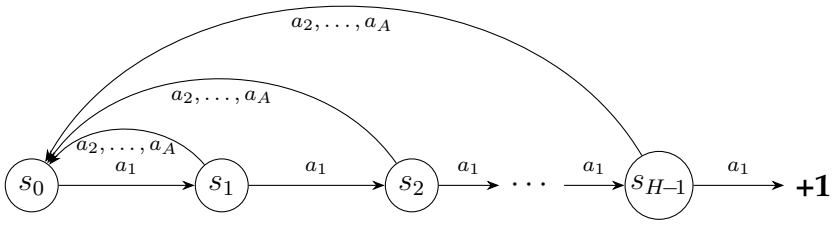
An algorithm achieves *sublinear regret* if $R_K = o(K)$, meaning the average per-episode suboptimality $R_K/K \to 0$.

Regret vs. PAC guarantees. If $R_K \le f(K)$ for some sublinear function f, then by an averaging argument, $\min_{k<K}[V_0^\star(s_0) - V_0^{\pi^k}(s_0)] \le f(K)/K$. So a regret bound of $\widetilde{O}(\sqrt{K})$ implies that after $K = \widetilde{O}(1/\varepsilon^2)$ episodes, at least one policy π^k is ε-optimal. This is an *online-to-batch* conversion that connects regret bounds to sample complexity (PAC) guarantees. Note the contrast with the generative model setting studied in earlier lectures: there, the agent could query any (s, a) pair at will, whereas here data collection is sequential and the agent cannot directly choose which states to visit.

Why Exploration Must Be Strategic

A natural baseline is to explore uniformly at random and then exploit. The following example shows this can be catastrophically inefficient.

The chain MDP. Consider the MDP depicted below with H states $s_0, s_1, \ldots, s_{H-1}$ arranged in a chain, with A actions at each state. At each state s_h ($h = 0, \ldots, H-2$), exactly one "correct" action a_1 transitions to s_{h+1}, while all other $A-1$ actions lead back to s_0. The reward is $r_{H-1}(s_{H-1}, a_1) = 1$ and zero everywhere else.



Lemma 1 (Random exploration is exponentially slow)**.** *A uniformly random policy collects nonzero reward with probability* $(1/A)^H$*. To achieve* $\Omega(1)$ *total reward, it needs* $\Omega(A^H)$ *episodes in expectation.*

Proof. Under a uniformly random policy, at each step h the correct action is chosen with probability $1/A$. The agent must choose correctly at all H steps: $H-1$ correct actions to traverse from s_0 to s_{H-1}, plus one correct action at s_{H-1} to collect the reward. The probability of success in a single episode is $(1/A)^H$, so the expected number of episodes needed is A^H. □

The need for strategic exploration. The chain MDP shows that uniform exploration can require exponentially many episodes. In contrast, a strategic agent can solve this MDP in $O(H)$ episodes: in the first episode, it discovers that action a_1 at s_0 leads to s_1; in the second, it can plan to visit s_1 and discover that a_1 at s_1 leads to s_2; and so on. The key insight is that the agent must *plan* to reach informative states, not just hope to stumble upon them. The *optimism in the face of uncertainty* (OFU) principle achieves this: by assuming the best case for unknown transitions, the optimistic agent is naturally drawn to explore under-visited states.

The Optimism Principle: Warm-up via Multi-Armed Bandits

Before tackling MDPs, let us review the optimism principle in the simplest possible setting: the K-armed bandit problem. There are K arms; pulling arm a yields a stochastic reward $r \in [0,1]$ with mean μ_a. Over T rounds, the learner selects arms $a_0, a_1, \ldots, a_{T-1}$ and observes rewards. The regret is $R_T = T\mu_\star - \sum_{t=0}^{T-1} \mu_{a_t}$, where $\mu_\star = \max_a \mu_a$.

The *Upper Confidence Bound* (UCB) algorithm (Auer et al., 2002) selects the arm maximizing the empirical mean plus an exploration bonus:

$$a_t = \operatorname*{argmax}_{a \in [K]} \left[\widehat{\mu}_t(a) + \sqrt{\frac{2\ln(2TK/\delta)}{N_t(a)}} \right],$$

where $\widehat{\mu}_t(a)$ is the empirical mean of arm a from the first $N_t(a)$ pulls, and the square root

term is the *exploration bonus* (or confidence radius).

Theorem 2 (UCB regret (Auer et al., 2002), stated without proof)**.** *With probability at least* $1-\delta$*, the UCB algorithm achieves*

$$R_T \le 2K + 8\sqrt{KT\ln(2TK/\delta)} = \widetilde{O}(\sqrt{KT}).$$

Three ingredients of the UCB analysis. The UCB proof relies on three components that will generalize to MDPs:

1. **Concentration**: By Hoeffding's inequality, $|\widehat{\mu}_t(a) - \mu_a| \le \sqrt{2\ln(2TK/\delta)/N_t(a)}$ simultaneously for all (t,a) with high probability. This ensures the bonus maintains a valid *upper confidence bound*.
2. **Optimism**: Since the bonus covers the estimation error, we have $\widehat{\mu}_t(a) + \text{bonus}_t(a) \ge \mu_a$ for all arms. In particular, $\text{UCB}_t(a^\star) \ge \mu_\star$, so the chosen arm's UCB is at least as large as $\mu_\star$.
3. **Shrinking bonuses + counting**: The per-round regret is at most $2\cdot\text{bonus}_t(a_t) \propto 1/\sqrt{N_t(a_t)}$. The sum $\sum_t 1/\sqrt{N_t(a_t)} \le 2\sqrt{KT}$ by a counting argument ($\sum_{i=1}^n 1/\sqrt{i} \le 2\sqrt{n}$ and Cauchy–Schwarz).

The challenge in extending to MDPs: the "arm" is a policy (a sequence of decisions), the "reward" depends on the entire trajectory, and the agent cannot directly choose which states to visit—it must plan to reach them.

UCBVI: UCB Value Iteration

We now present the UCBVI algorithm for the tabular episodic setting, following Azar et al. (2017) and Agarwal et al. (2021).

Algorithm Description

Counts and empirical model. For each stage $h \in \{0,\dots,H-1\}$ and $(s,a) \in \mathcal{S}\times\mathcal{A}$, define the pre-episode-k visitation counts:

$$N_h^k(s,a) := \sum_{i=0}^{k-1} \mathbb{1}\{(s_h^i, a_h^i) = (s,a)\}, \qquad N_h^k(s,a,s') := \sum_{i=0}^{k-1} \mathbb{1}\{(s_h^i, a_h^i, s_{h+1}^i) = (s,a,s')\}.$$

The empirical transition model is

$$\widehat{P}_h^k(s' \mid s,a) := \begin{cases} N_h^k(s,a,s')/N_h^k(s,a) & \text{if } N_h^k(s,a) > 0, \\ \text{any distribution on } \mathcal{S} & \text{if } N_h^k(s,a) = 0. \end{cases}$$

Algorithm 1 UCBVI (episodic, tabular) (Azar et al., 2017)

Require: Failure probability $\delta \in (0,1)$, horizon H, episodes K
1: Initialize $N_h^0(s,a) \leftarrow 0$ and $N_h^0(s,a,s') \leftarrow 0$ for all h, s, a, s'
2: **for** $k = 0, 1, \ldots, K-1$ **do**
3: Form $\widehat{P}_h^k(\cdot \mid s,a)$ from counts ▷ *empirical model*
4: Set bonuses $b_h^k(s,a)$ by (1) ▷ *exploration bonus*
5: Compute $(\widehat{V}_h^k, \widehat{Q}_h^k, \pi_h^k)_{h=0}^{H-1}$ by (2) ▷ *optimistic backward induction*
6: Execute π^k: set $s_0^k = s_0$; for $h = 0, \ldots, H-1$, take $a_h^k = \pi_h^k(s_h^k)$, sample $s_{h+1}^k \sim P_h^\star(\cdot \mid s_h^k, a_h^k)$
7: Update counts: $N_h^{k+1}(\cdot) \leftarrow N_h^k(\cdot)$ then increment the visited $(s_h^k, a_h^k, s_{h+1}^k)$
8: **end for**

When $N_h^k(s,a) = 0$, the model is arbitrary; the exploration bonus defined next ensures that such state-action pairs are treated as highly uncertain.

Exploration bonus. Fix a failure probability $\delta \in (0,1)$ and define the log factor $L := \ln(2SAHK/\delta)$. For each episode k, stage h, and (s,a), define

$$b_h^k(s,a) \;:=\; \min\left\{H-h,\;\; 2H\sqrt{\frac{L}{N_h^k(s,a)+1}}\right\}. \tag{1}$$

The bonus is large when (s,a) has been visited few times and decays as $1/\sqrt{N_h^k(s,a)}$ as more data are collected. The clipping by $H-h$ matches the maximum possible remaining return from stage h.

Optimistic planning. Given the empirical model $\{\widehat{P}_h^k\}$ and bonuses $\{b_h^k\}$, UCBVI performs a backward dynamic program. Initialize $\widehat{V}_H^k(\cdot) \equiv 0$ and for $h = H-1, \ldots, 0$ compute

$$\widehat{Q}_h^k(s,a) := r_h(s,a) + b_h^k(s,a) + \widehat{P}_h^k \widehat{V}_{h+1}^k(s,a), \qquad \widehat{V}_h^k(s) := \min\left\{H-h,\;\; \max_{a\in\mathcal{A}} \widehat{Q}_h^k(s,a)\right\}. \tag{2}$$

The greedy policy is $\pi_h^k(s) := \operatorname{argmax}_{a\in\mathcal{A}} \widehat{Q}_h^k(s,a)$.

Optimistic Bellman equation:

$$\widehat{Q}_h^k(s,a) = r_h(s,a) + \underbrace{b_h^k(s,a)}_{\text{exploration bonus}} + \underbrace{\widehat{P}_h^k \widehat{V}_{h+1}^k(s,a)}_{\text{empirical Bellman backup}}.$$

Interpretation. UCBVI combines three ideas: (1) *certainty equivalence* (use empirical transitions $\widehat{P}_h^k$ in place of $P_h^\star$), (2) *exploration bonuses* (inflate Q-values for under-visited state-actions), and (3) *online adaptation* (update the model after each episode). The bonus ensures that the agent is "optimistic"—it overestimates the value of under-explored state-action pairs, and is thus naturally drawn to explore them. Unlike the generative model setting, there is no simulator; the agent must sequentially visit states.

Main Result

Theorem 3 (UCBVI regret bound, simple analysis)**.** *Run UCBVI (Algorithm 1) with bonuses* (1)*. For any $\delta \in (0,1)$, letting $L = \ln(2SAHK/\delta)$,*

$$R_K \;\le\; O\Big(H^2 S\sqrt{AK \cdot L}\Big) + \delta KH.$$

In particular, setting $\delta = 1/(KH)$ yields

$$R_K \;=\; \widetilde{O}\Big(H^2 S\sqrt{AK}\Big)\,.$$

UCBVI regret (simple analysis):

$$R_K = \widetilde{O}\Big(H^2 S\sqrt{AK}\Big)\,.$$

Interpretation and comparison.

- The regret is sublinear in K: the average per-episode suboptimality is $\widetilde{O}(H^2 S\sqrt{A/K}) \to 0$.
- The trivial bound is $R_K \le KH$ (since each episode contributes at most H regret). Theorem 3 becomes nontrivial once $K \gtrsim H^2 S^2 A$ (up to logs).
- Compared to random exploration, which needs $\Omega(A^H)$ episodes for the chain MDP, UCBVI needs only $\widetilde{O}(H^4 S^2 A/\varepsilon^2)$ episodes for an ε-optimal average policy.
- The minimax lower bound is $\Omega(H^{3/2}\sqrt{SAK})$. The simple analysis pays an extra $\sqrt{S}$ factor; the refined Bernstein analysis (Section) achieves $\widetilde{O}(H^2\sqrt{SAK})$, which is near-optimal up to a $\sqrt{H}$ factor.

Analysis of UCBVI

Proof Overview

The proof of Theorem 3 has four main components:

1. **Concentration (Lemmas 4 and 5):** The empirical model $\widehat{P}_h^k$ is close to the true model $P_h^\star$ with high probability, uniformly over all (k, h, s, a).
2. **Optimism (Lemma 6):** The optimistic value function upper-bounds the true optimal value: $\widehat{V}_h^k(s) \geq V_h^\star(s)$ for all (k, h, s).
3. **One-episode decomposition (Lemma 7):** The per-episode regret telescopes into a sum of bonuses and model errors along the trajectory.
4. **Counting lemma (Lemma 8):** The total exploration cost $\sum_{k,h} 1/\sqrt{N_h^k(s_h^k, a_h^k) + 1}$ is bounded by $O(H\sqrt{SAK})$.

The canonical four-step template. This four-step proof structure—concentration, optimism, decomposition, counting—is the canonical template for regret analyses of optimistic exploration algorithms. The same structure applies to virtually all UCB-style algorithms in RL, including the linear MDP case we study later. The key insight is that optimism converts the exploration problem into a "pay for the bonus" problem: the agent's regret is bounded by what it "spends" on exploration bonuses, and the bonuses shrink as data accumulates.

Step 1: Concentration

We first establish that the empirical transition model concentrates around the true model, uniformly over all episodes.

Lemma 4 (Uniform ℓ_1 transition concentration)**.** *Fix $\delta \in (0, 1)$ and let $L = \ln(2SAHK/\delta)$. With probability at least $1 - \delta/2$, simultaneously for all $h \in [H]$, all $(s, a) \in \mathcal{S} \times \mathcal{A}$, and all episodes k with $N_h^k(s, a) \geq 1$,*

$$\left\|\widehat{P}_h^k(\cdot \mid s, a) - P_h^\star(\cdot \mid s, a)\right\|_1 \;\leq\; 2\sqrt{\frac{SL}{N_h^k(s, a)}}.$$

Consequently, for any (possibly data-dependent) function $f : \mathcal{S} \to [0, H]$,

$$\left|(\widehat{P}_h^k - P_h^\star)^\top f(s, a)\right| \;\leq\; H \cdot \left\|\widehat{P}_h^k - P_h^\star\right\|_1 \;\leq\; 2H\sqrt{\frac{SL}{N_h^k(s, a)}}.$$

Proof. Fix (h, s, a) and consider the empirical distribution after exactly n i.i.d. samples from $P_h^\star(\cdot \mid s, a)$; denote it by $\widehat{P}_{h,n}(\cdot \mid s, a)$. By a standard bound on the ℓ_1 distance between an empirical distribution and the true distribution over a support of size S (using Hoeffding's inequality applied to each next-state indicator followed by a union bound), with probability

at least $1-\delta'$,

$$\left\|\widehat{P}_{h,n}(\cdot \mid s,a) - P_h^\star(\cdot \mid s,a)\right\|_1 \;\le\; 2\sqrt{\frac{S\ln(1/\delta')}{n}}.$$

Choose $\delta' = \delta/(2SAHK)$ and take a union bound over all $(h,s,a) \in [H]\times\mathcal{S}\times\mathcal{A}$ and all sample sizes $n \in \{1,\dots,K\}$. This gives a bound that holds simultaneously for every fixed n. Now, although the online count $N_h^k(s,a)$ is random (it depends on the trajectory history), it always takes a value in $\{0,1,\dots,K\}$. When $N_h^k(s,a) = n$, the online empirical distribution $\widehat{P}_h^k(\cdot \mid s,a)$ is computed from the same n i.i.d. samples, so $\widehat{P}_h^k = \widehat{P}_{h,n}$. Since the bound already covers every $n \in \{1,\dots,K\}$, it holds for the random count $N_h^k(s,a)$ as well, simultaneously for all episodes k.

The final display follows from Hölder's inequality: $|(\widehat{P}-P)^\top f| \le \|\widehat{P}-P\|_1 \cdot \|f\|_\infty \le H\|\widehat{P}-P\|_1$. □

We also need a sharper bound for the deterministic optimal value function $V_{h+1}^\star$ (to prove optimism without paying a $\sqrt{S}$ factor at this step).

Lemma 5 (Model error against $V^\star$)**.** *Fix $\delta \in (0,1)$ and let $L = \ln(2SAHK/\delta)$. With probability at least $1-\delta/2$, simultaneously for all $h \in [H]$, all (s,a), and all episodes k with $N_h^k(s,a) \ge 1$,*

$$\left|(\widehat{P}_h^k - P_h^\star)^\top V_{h+1}^\star(s,a)\right| \;\le\; H\sqrt{\frac{L}{N_h^k(s,a)}}.$$

Proof. Fix (h,s,a) and consider n i.i.d. samples $s'_1,\dots,s'_n \sim P_h^\star(\cdot \mid s,a)$. Let $Y_i := V_{h+1}^\star(s'_i) \in [0,H]$, so $\mathbb{E}[Y_i] = (P_h^\star)^\top V_{h+1}^\star(s,a)$. By Hoeffding's inequality,

$$\Pr\left(\left|\frac{1}{n}\sum_{i=1}^n Y_i - \mathbb{E}[Y_i]\right| \ge \varepsilon\right) \;\le\; 2\exp\left(-\frac{2n\varepsilon^2}{H^2}\right).$$

Setting $\varepsilon = H\sqrt{L/n}$ makes the right-hand side at most $\delta/(2SAHK)$ (using $L \ge \ln 2$). A union bound over all $(h,s,a) \in [H]\times\mathcal{S}\times\mathcal{A}$ and all $n \in \{1,\dots,K\}$ gives a bound for every fixed sample size. Since $N_h^k(s,a) \in \{1,\dots,K\}$ and the online samples coincide with the first $N_h^k(s,a)$ i.i.d. draws (as in the proof of Lemma 4), the bound holds for the random count as well. □

Let $\mathcal{E}_{\text{model}}$ denote the intersection of the events in Lemmas 4 and 5. Then $\Pr(\mathcal{E}_{\text{model}}) \ge 1-\delta$.

The source of the extra $\sqrt{S}$. Lemma 4 bounds the model error against *any* bounded function f by $O(H\sqrt{SL/N})$, paying a $\sqrt{S}$ factor through the ℓ_1 bound. Lemma 5 avoids this $\sqrt{S}$ for the specific function $V^\star_{h+1}$, but we cannot use Lemma 5 for the algorithm's own (random) value function $\widehat{V}^k_{h+1}$. This is why the simple analysis incurs an extra $\sqrt{S}$: it uses the ℓ_1 bound to handle the model error against $\widehat{V}^k_{h+1}$. The refined Bernstein analysis (Section) avoids this by using variance-dependent bounds.

Step 2: Optimism

Lemma 6 (Optimism). *On the event $\mathcal{E}_{\text{model}}$, for every episode k, every stage h, and every state s:*

$$\widehat{V}^k_h(s) \;\geq\; V^\star_h(s).$$

In particular, $\widehat{V}^k_0(s_0) \geq V^\star_0(s_0)$ for all k.

Proof. We prove by backward induction on h.

Base case: At $h = H$, $\widehat{V}^k_H \equiv 0 = V^\star_H$.

Inductive step: Assume $\widehat{V}^k_{h+1}(s) \geq V^\star_{h+1}(s)$ for all s. Fix a state s at stage h.

If $\widehat{V}^k_h(s) = H - h$, then $\widehat{V}^k_h(s) \geq V^\star_h(s)$ holds since $V^\star_h(s) \leq H - h$.

Otherwise, $\widehat{V}^k_h(s) = \max_a \widehat{Q}^k_h(s,a)$. Let $a \in \mathcal{A}$ be arbitrary. Using Lemma 5 and the induction hypothesis:

$$\begin{aligned}
Q^\star_h(s,a) &= r_h(s,a) + (P^\star_h)^\top V^\star_{h+1}(s,a) \\
&\leq r_h(s,a) + (\widehat{P}^k_h)^\top V^\star_{h+1}(s,a) + H\sqrt{\frac{L}{N^k_h(s,a)}} \\
&\leq r_h(s,a) + (\widehat{P}^k_h)^\top \widehat{V}^k_{h+1}(s,a) + H\sqrt{\frac{L}{N^k_h(s,a)}} \\
&\leq r_h(s,a) + (\widehat{P}^k_h)^\top \widehat{V}^k_{h+1}(s,a) + b^k_h(s,a) \;=\; \widehat{Q}^k_h(s,a),
\end{aligned}$$

where the second line uses Lemma 5 on the event $\mathcal{E}_{\text{model}}$, the third line uses $\widehat{V}^k_{h+1} \geq V^\star_{h+1}$ (induction) and the fact that $\widehat{P}^k_h$ is a probability distribution (so applying it to a pointwise larger function gives a larger result), and the fourth line uses $b^k_h(s,a) \geq H\sqrt{L/N^k_h(s,a)}$ from the bonus definition (1) (since $N^k_h(s,a) + 1 \leq 2N^k_h(s,a)$ when $N^k_h(s,a) \geq 1$; when $N^k_h(s,a) = 0$, the bonus $b^k_h(s,a) = H - h \geq Q^\star_h(s,a) - r_h(s,a)$ suffices directly).

Taking $\max_a$ over both sides gives $V^\star_h(s) = \max_a Q^\star_h(s,a) \leq \max_a \widehat{Q}^k_h(s,a) = \widehat{V}^k_h(s)$. □

Structure of the optimism proof. The backward induction structure of the optimism proof mirrors the backward induction of the algorithm itself. This is not a coincidence—it is the key structural feature that makes value-function-based exploration tractable. The proof requires two ingredients: (i) the bonus is large enough to cover the model error against $V^\star_{h+1}$, and (ii) $\widehat{P}^k_h$ applied to a pointwise larger function gives a larger result (since $\widehat{P}^k_h$ is a valid probability distribution). Together, these ensure that adding the bonus to the empirical Bellman backup yields an *upper* bound on $Q^\star_h$.

Step 3: One-Episode Regret Decomposition

Lemma 7 (One-episode decomposition)**.** *For any episode k, the policy π^k returned by UCBVI satisfies*

$$\widehat{V}^k_0(s_0) - V^{\pi^k}_0(s_0) \;\le\; \sum_{h=0}^{H-1} \mathbb{E}_{(s_h,a_h)\sim d^{\pi^k}_h}\left[b^k_h(s_h,a_h) + \left|(\widehat{P}^k_h - P^\star_h)^\top \widehat{V}^k_{h+1}(s_h,a_h)\right|\right].$$

Proof. Fix episode k and abbreviate $\pi = \pi^k$, $\widehat{V}_h = \widehat{V}^k_h$, $\widehat{P}_h = \widehat{P}^k_h$, $b_h = b^k_h$. Define the gap $\Delta_h(s) \coloneqq \widehat{V}_h(s) - V^\pi_h(s)$, with $\Delta_H \equiv 0$.

For any state s at stage h, since $\pi_h(s) \in \operatorname{argmax}_a \widehat{Q}_h(s,a)$ and $\widehat{V}_h(s) = \min\{H-h, \max_a \widehat{Q}_h(s,a)\} \le \widehat{Q}_h(s,\pi_h(s))$, we have

$$\widehat{V}_h(s) \le r_h(s,\pi_h(s)) + b_h(s,\pi_h(s)) + \widehat{P}_h^\top \widehat{V}_{h+1}(s,\pi_h(s)).$$

On the other hand, $V^\pi_h(s) = r_h(s,\pi_h(s)) + (P^\star_h)^\top V^\pi_{h+1}(s,\pi_h(s))$.

Subtracting:

$$\Delta_h(s) \;\le\; b_h(s,\pi_h(s)) + \left(\widehat{P}_h - P^\star_h\right)^\top \widehat{V}_{h+1}(s,\pi_h(s)) + (P^\star_h)^\top \Delta_{h+1}(s,\pi_h(s)).$$

Taking expectations under the trajectory distribution induced by π in the true MDP and unrolling (using the tower property) from $h=0$ to $H-1$, with $\Delta_H \equiv 0$, yields the stated bound for $\Delta_0(s_0) = \widehat{V}_0(s_0) - V^\pi_0(s_0)$. □

Combining with optimism ($V^\star_0(s_0) \le \widehat{V}^k_0(s_0)$ from Lemma 6):

One-episode regret decomposition: On $\mathcal{E}_{\text{model}}$,

$$V^\star_0(s_0) - V^{\pi^k}_0(s_0) \;\le\; \sum_{h=0}^{H-1} \mathbb{E}_{(s_h,a_h)\sim d^{\pi^k}_h}\left[b^k_h(s_h,a_h) + \left|(\widehat{P}^k_h - P^\star_h)^\top \widehat{V}^k_{h+1}(s_h,a_h)\right|\right].$$

Regret is controlled by bonuses along the trajectory. The decomposition shows that the per-episode regret is at most the sum of exploration bonuses and model errors encountered along the trajectory induced by π^k. On the event $\mathcal{E}_{\text{model}}$, the model error $|(\widehat{P}_h^k - P_h^\star)^\top \widehat{V}_{h+1}^k|$ can be bounded via Lemma 4 by $O(H\sqrt{SL/N_h^k})$, which has the same scaling as the bonus. The key remaining question is: how large is the sum $\sum_k \sum_h b_h^k(s_h^k, a_h^k)$ across all episodes?

Step 4: The Counting Lemma

Lemma 8 (Counting / potential bound (Agarwal et al., 2021, Lemma 6.6))**.** *For any sequence of K trajectories $\{(s_h^k, a_h^k)\}_{h=0}^{H-1}$, we have*

$$\sum_{k=0}^{K-1} \sum_{h=0}^{H-1} \frac{1}{\sqrt{N_h^k(s_h^k, a_h^k) + 1}} \;\le\; 2H\sqrt{SAK}.$$

Proof. Fix a stage h and a state–action pair (s, a). Let $n \coloneqq N_h^K(s, a)$ be the total number of visits to (s, a) at stage h over all K episodes. Each time (s, a) is visited, the denominator takes values $\sqrt{1}, \sqrt{2}, \ldots, \sqrt{n}$ (since N_h^k is the count *before* episode k, the relevant denominator values are $\sqrt{0+1}, \sqrt{1+1}, \ldots, \sqrt{(n-1)+1}$). Therefore:

$$\sum_{k:\,(s_h^k, a_h^k)=(s,a)} \frac{1}{\sqrt{N_h^k(s, a) + 1}} \;\le\; \sum_{i=1}^{n} \frac{1}{\sqrt{i}} \;\le\; 2\sqrt{n} = 2\sqrt{N_h^K(s, a)}.$$

Summing over all $(s, a) \in \mathcal{S} \times \mathcal{A}$ and applying Cauchy–Schwarz:

$$\sum_{s,a} 2\sqrt{N_h^K(s, a)} \;\le\; 2\sqrt{SA \sum_{s,a} N_h^K(s, a)} \;=\; 2\sqrt{SA \cdot K},$$

where we used $\sum_{s,a} N_h^K(s, a) = K$ (each episode visits exactly one (s, a) at stage h).

Finally, summing over $h = 0, \ldots, H-1$ gives $2H\sqrt{SAK}$. □

Counting lemma:

$$\sum_{k=0}^{K-1} \sum_{h=0}^{H-1} \frac{1}{\sqrt{N_h^k(s_h^k, a_h^k) + 1}} \;\le\; 2H\sqrt{SAK}.$$

The paying-for-exploration principle. The counting lemma quantifies the total "exploration cost." Early on, bonuses are large (when N is small), but as the agent accumulates data, the bonuses shrink. The total cost is controlled by the harmonic-like sum $\sum_{i=1}^{n} 1/\sqrt{i} \le 2\sqrt{n}$. The Cauchy–Schwarz step produces the $\sqrt{SA}$ factor: spreading K visits across SA state-action pairs costs $\sqrt{SA}$ via Jensen's inequality (the sum $\sum \sqrt{n_i}$ is maximized when all n_i are equal).

Completing the Proof of Theorem 3

Proof of Theorem 3. Fix $\delta \in (0,1)$ and let $\mathcal{E}_{\text{model}}$ be as above, with $\Pr(\mathcal{E}_{\text{model}}) \ge 1-\delta$.

Step 1: Reduce regret to the optimistic gap. On $\mathcal{E}_{\text{model}}$, optimism (Lemma 6) implies

$$V_0^\star(s_0) - V_0^{\pi^k}(s_0) \;\le\; \widehat{V}_0^k(s_0) - V_0^{\pi^k}(s_0).$$

Step 2: Decompose and bound the model error. By Lemma 7:

$$\widehat{V}_0^k(s_0) - V_0^{\pi^k}(s_0) \;\le\; \sum_{h=0}^{H-1} \mathbb{E}_{d_h^{\pi^k}}\left[b_h^k(s_h,a_h) + \left|(\widehat{P}_h^k - P_h^\star)^\top \widehat{V}_{h+1}^k(s_h,a_h)\right| \right].$$

Since $\|\widehat{V}_{h+1}^k\|_\infty \le H$, Lemma 4 gives (for $N_h^k(s_h,a_h) \ge 1$)

$$\left|(\widehat{P}_h^k - P_h^\star)^\top \widehat{V}_{h+1}^k(s_h,a_h)\right| \;\le\; 2H\sqrt{\frac{SL}{N_h^k(s_h,a_h)}} \;\le\; 2\sqrt{2}\,H\sqrt{\frac{SL}{N_h^k(s_h,a_h)+1}},$$

where we used $N+1 \le 2N$ for $N \ge 1$. When $N_h^k(s_h,a_h) = 0$, we trivially have $|(\widehat{P}_h^k - P_h^\star)^\top \widehat{V}_{h+1}^k| \le H$, which is also bounded by $2\sqrt{2}\,H\sqrt{SL/(0+1)}$ since $H \le 2\sqrt{2}\,H\sqrt{SL}$ (using $S \ge 1$ and $L \ge \ln 2$).

Similarly, the bonus satisfies $b_h^k(s_h,a_h) \le 2H\sqrt{L/(N_h^k(s_h,a_h)+1)} \le 2H\sqrt{SL/(N_h^k(s_h,a_h)+1)}$.

Step 3: Sum over episodes. Combining, on $\mathcal{E}_{\text{model}}$:

$$V_0^\star(s_0) - V_0^{\pi^k}(s_0) \;\le\; (2+2\sqrt{2})\,H\sqrt{SL}\cdot \mathbb{E}\left[\sum_{h=0}^{H-1} \frac{1}{\sqrt{N_h^k(s_h^k,a_h^k)+1}} \,\middle|\, \mathcal{H}_{<k}\right].$$

Summing over k and taking total expectation, then applying the counting lemma (Lemma 8):

$$\mathbb{E}\left[\mathbb{1}\{\mathcal{E}_{\text{model}}\}\sum_{k=0}^{K-1}\left(V_0^\star(s_0) - V_0^{\pi^k}(s_0)\right)\right] \;\le\; (2+2\sqrt{2})\,H\sqrt{SL}\cdot 2H\sqrt{SAK} \;=\; O\Big(H^2 S\sqrt{AK\cdot L}\Big).$$

Step 4: Handle the failure event. On $\overline{\mathcal{E}}_{\text{model}}$ (probability $\leq \delta$), use the trivial bound $0 \leq V_0^\star(s_0) - V_0^{\pi^k}(s_0) \leq H$:

$$\mathbb{E}\left[\mathbb{1}\{\overline{\mathcal{E}}_{\text{model}}\} \sum_{k=0}^{K-1} \left(V_0^\star(s_0) - V_0^{\pi^k}(s_0)\right)\right] \;\leq\; \delta \cdot KH.$$

Combining gives $R_K \leq O(H^2 S\sqrt{AK \cdot L}) + \delta KH$. □

Refined Analysis via Bernstein Bonuses

This section follows Agarwal et al. (2021, Section 6.5, especially Lemmas 6.8–6.11 and Theorem 6.7).

Motivation: Removing the Extra $\sqrt{S}$

The simple analysis bounds the model error $|(\widehat{P}_h^k - P_h^\star)^\top \widehat{V}_{h+1}^k|$ using the crude ℓ_1 bound $\|\widehat{P} - P^\star\|_1 \cdot \|\widehat{V}\|_\infty$, which introduces a $\sqrt{S}$ factor. However, if the *variance* of $\widehat{V}_{h+1}^k(s')$ under $P_h^\star(\cdot \mid s, a)$ is small, the model error can be much smaller than the worst-case ℓ_1 bound suggests. Bernstein's inequality (from Lecture 5) gives tighter bounds that depend on variance rather than range.

Bernstein-Style Transition Bound

The refined argument begins with a pointwise empirical Bernstein bound for each next-state probability. This is the step that replaces the crude ℓ_1 control from the simple analysis.

Lemma 9 (Pointwise empirical Bernstein bound)**.** *Fix $\delta \in (0,1)$ and let $L = \ln(2S^2AHK/\delta)$. With probability at least $1-\delta$, simultaneously for all $h \in \{0,\ldots,H-1\}$, all $(s,a) \in \mathcal{S} \times \mathcal{A}$, all $s' \in \mathcal{S}$, and all episodes $k \in \{0,\ldots,K-1\}$,*

$$\widehat{P}_h^k(s' \mid s,a) - P_h^\star(s' \mid s,a) \;\leq\; 2\sqrt{\frac{P_h^\star(s' \mid s,a)\, L}{N_h^k(s,a)+1}} \;+\; \frac{4L}{N_h^k(s,a)+1}.$$

Let $\mathcal{E}_{\text{bern}}$ denote the event in Lemma 9. On this event, we obtain the following self-bounding inequality:

Lemma 10 (Self-bounding transition error)**.** *On the event $\mathcal{E}_{\text{bern}}$, for all h, k, s, a and all functions*

$f : \mathcal{S} \to [0, H]$,

$$\big((\widehat{P}_h^k - P_h^\star)^\top f\big)(s,a) \;\le\; \frac{\big((P_h^\star)^\top f\big)(s,a)}{H} \;+\; c_h^k(s,a), \qquad c_h^k(s,a) := \frac{c\,H^2 SL}{N_h^k(s,a)+1},$$

for a universal constant $c > 0$.

The crucial feature is that the leading term $(P_h^\star)^\top f/H$ is *self-bounding*: it depends on the expected value of f under the true transition, not on the worst-case range. This is the key mechanism that removes the extra $\sqrt{S}$ from the leading regret term.

Improved Regret Bound

Using Lemma 10, one can establish a *recursive control* of the optimistic gap $\Delta_h^k(s) := \widehat{V}_h^k(s) - V_h^\star(s) \ge 0$:

Lemma 11 (Gap recursion). *Assume* $\mathcal{E}_{\mathrm{model}}$ *holds (so* $\widehat{V}^k \ge V^\star$ *by Lemma 6) and Lemma 10 holds. Then for each episode* k*, stage* h*, and state* s*, letting* $a = \pi_h^k(s)$*:*

$$\Delta_h^k(s) \;\le\; 2b_h^k(s,a) + c_h^k(s,a) + \left(1 + \frac{1}{H}\right) \mathbb{E}_{s' \sim P_h^\star(\cdot|s,a)}\big[\Delta_{h+1}^k(s')\big]\,.$$

Consequently,

$$\begin{aligned}\Delta_h^k(s) \;&\le\; \mathbb{E}\left[\sum_{\tau=h}^{H-1} \left(1+\frac{1}{H}\right)^{\tau-h} \big(2b_\tau^k(s_\tau,a_\tau) + c_\tau^k(s_\tau,a_\tau)\big) \;\middle|\; s_h = s, \pi^k\right] \\ &\le\; e \cdot \mathbb{E}\left[\sum_{\tau=h}^{H-1} \big(2b_\tau^k(s_\tau,a_\tau) + c_\tau^k(s_\tau,a_\tau)\big) \;\middle|\; s_h = s, \pi^k\right].\end{aligned}$$

The recursion yields the following one-episode bound, which is the refined analogue of Lemma 7.

Lemma 12 (Per-episode refined bound). *Assume* $\mathcal{E}_{\mathrm{model}}$ *and Lemmas 10–11 hold. Then for every episode* k*,*

$$V_0^\star(s_0) - V_0^{\pi^k}(s_0) \;\le\; e \sum_{h=0}^{H-1} \mathbb{E}_{(s_h,a_h)\sim d_h^{\pi^k}}\big[2b_h^k(s_h,a_h) + c_h^k(s_h,a_h)\big]\,.$$

This leads to:

Theorem 13 (UCBVI regret bound, refined analysis). *Run UCBVI (Algorithm 1) with*

bonuses (1). *There exists a universal constant* $C > 0$ *such that for any* $\delta \in (0,1)$*, letting* $L = \ln(2S^2AHK/\delta)$*,*

$$R_K \;\le\; C\Big(H^2\sqrt{SAK\cdot L} + H^3S^2A\cdot L\cdot \ln(1+K)\Big) + \delta KH.$$

In particular, setting $\delta = 1/(KH)$ *yields* $R_K = \widetilde{O}(H^2\sqrt{SAK} + H^3S^2A)$*.*

UCBVI regret (refined analysis):

$$R_K = \widetilde{O}\Big(H^2\sqrt{SAK}\Big),$$

where the leading term is $\widetilde{O}(H^2\sqrt{SAK})$ and the lower-order term $\widetilde{O}(H^3S^2A)$ is negligible for $K \gg H^2S^3A$.

Proof sketch. Let $\mathcal{E} := \mathcal{E}_{\text{model}} \cap \mathcal{E}_{\text{bern}}$. Then $\Pr(\mathcal{E}) \ge 1-\delta$ after the standard $\delta/2$ split between the two concentration events. The proof follows the same four-step template as the simple analysis. The key differences are:

(i) Variance-dependent model error. Instead of bounding the model error via ℓ_1, Lemma 10 bounds the *one-sided* error $(\widehat{P}_h^k - P_h^\star)^\top f$ by $(P_h^\star)^\top f/H + c_h^k$. The first term is self-bounding and is absorbed into the recursion of Lemma 11, producing the $(1+1/H)$ factor; unrolling gives at most an e loss.

(ii) Per-episode regret bound. On $\mathcal{E}$, Lemma 12 gives

$$V_0^\star(s_0) - V_0^{\pi^k}(s_0) \;\le\; e\sum_{h=0}^{H-1}\mathbb{E}_{(s_h,a_h)\sim d_h^{\pi^k}}\left[2b_h^k(s_h,a_h) + c_h^k(s_h,a_h)\right].$$

(iii) Summation. Summing over episodes and using total expectation,

$$\mathbb{E}\left[\mathbb{1}\{\mathcal{E}\}\sum_{k=0}^{K-1}\big(V_0^\star(s_0) - V_0^{\pi^k}(s_0)\big)\right] \;\le\; e\,\mathbb{E}\left[\sum_{k=0}^{K-1}\sum_{h=0}^{H-1}\big(2b_h^k(s_h^k,a_h^k) + c_h^k(s_h^k,a_h^k)\big)\right].$$

Using $b_h^k(s,a) \le 2H\sqrt{L/(N_h^k(s,a)+1)}$ and the counting lemma (Lemma 8) gives

$$\sum_{k,h} 2b_h^k(s_h^k,a_h^k) \;\le\; 4H\sqrt{L}\sum_{k,h}\frac{1}{\sqrt{N_h^k(s_h^k,a_h^k)+1}} \;\le\; 8H^2\sqrt{SAK\cdot L}.$$

The lower-order term satisfies

$$\sum_{k,h} c_h^k(s_h^k, a_h^k) \;=\; O(H^3 S^2 A \cdot L \cdot \ln(1+K))$$

by summing the harmonic series over visits to each (h, s, a). On $\overline{\mathcal{E}}$, the trivial bound $0 \le V_0^\star(s_0) - V_0^{\pi^k}(s_0) \le H$ contributes at most $\delta K H$. □

Near-minimax-optimal. The improved bound $\widetilde{O}(H^2\sqrt{SAK})$ matches the minimax lower bound $\Omega(H^{3/2}\sqrt{SAK})$ up to an $\sqrt{H}$ factor. Closing this gap requires more sophisticated algorithms and analyses; see the bibliographic notes in Agarwal et al. (2021). The Bernstein analysis pattern—using variance-dependent concentration to save factors—is a recurring theme in RL theory.

Linear MDP and Lin-UCBVI

We now extend the UCBVI framework to MDPs with *linear structure*, where the state and action spaces may be large (or even infinite), but the transition dynamics are linearly parameterized by a known feature map. The algorithm originates in Jin et al. (2020); the presentation and analysis here follow Agarwal et al. (2021, Section 7.4–7.5).

The Linear Transition Model

Definition 2 (Linear transition model)**.** *An episodic MDP $M = (\mathcal{S}, \mathcal{A}, H, \{P_h^\star\}, \{r_h\})$ satisfies the* linear transition model *with feature map $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ if there exist unknown signed measures $\mu_h^\star = (\mu_h^{\star(1)}, \ldots, \mu_h^{\star(d)})$ over $\mathcal{S}$ such that for all (s, a, h),*

$$P_h^\star(\cdot \mid s, a) = \langle \phi(s, a),\, \mu_h^\star(\cdot) \rangle,$$

and the right-hand side is a valid probability distribution on $\mathcal{S}$. We assume $\|\phi(s, a)\|_2 \le 1$ for all (s, a), and

$$\|(\mu_h^\star)^\top f\|_2 \le \sqrt{d}\, \|f\|_\infty$$

for all $f : \mathcal{S} \to \mathbb{R}$ and all h.

Linear transition model: The one-step transition kernel is linear in the features:

$$P_h^\star(\cdot \mid s, a) = \langle \phi(s, a), \mu_h^\star(\cdot) \rangle,$$

$$\|(\mu_h^\star)^\top f\|_2 \le \sqrt{d}\, \|f\|_\infty, \qquad r_h(s, a) \in [0, 1].$$

Lemma 14 (Linearization of $P_h^\star f$)**.** *Under the linear transition model, for any function $f : \mathcal{S} \to \mathbb{R}$,*

$$(P_h^\star f)(s,a) = \langle \phi(s,a),\, w_{h,f}^\star \rangle, \qquad \textit{where } w_{h,f}^\star \coloneqq (\mu_h^\star)^\top f \in \mathbb{R}^d.$$

Moreover, if $\|f\|_\infty \le H$, then $\|w_{h,f}^\star\|_2 \le H\sqrt{d}$.

Key properties of the linear transition model.

- **One-step linearization**: For *any* bounded function f (not just linear ones), the one-step conditional expectation $(P_h^\star f)(s,a)$ remains linear in $\phi(s,a)$. This is the key completeness property behind Lin-UCBVI: even though the value functions produced by max and clipping are nonlinear, the transition term fed into regression is still linear in the features.
- **Tabular as special case**: Setting $\phi(s,a) = e_{(s,a)} \in \mathbb{R}^{SA}$ (the one-hot encoding) recovers the tabular setting with $d = SA$. In this case, $\mu_h^\star(s') = P_h^\star(s' \mid \cdot, \cdot)$ and the feature dimension equals the number of state-action pairs.
- **Dimension replaces cardinality**: The regret of Lin-UCBVI will scale with the feature dimension d, not with $|\mathcal{S}|$ or $|\mathcal{A}|$. When $d \ll SA$, this gives a substantial improvement.

Algorithm: Lin-UCBVI

Ridge regression for the model. At episode k, for each stage h, we estimate the linear coefficient $w_{h,f}^\star$ for a given target function f via ridge regression. Define the regularized design matrix

$$\Lambda_h^k \;\coloneqq\; \lambda I + \sum_{i=0}^{k-1} \phi(s_h^i, a_h^i)\, \phi(s_h^i, a_h^i)^\mathsf{T}, \qquad \lambda \ge 1.$$

Given a target function $f : \mathcal{S} \to \mathbb{R}$ (in our case, $f = \widehat{V}_{h+1}^k$), the ridge regression estimate is

$$\widehat{w}_{h,f}^k \;\coloneqq\; (\Lambda_h^k)^{-1} \sum_{i=0}^{k-1} \phi(s_h^i, a_h^i)\, f(s_{h+1}^i),$$

so that the estimated Bellman backup is $(\widehat{P}_h^k f)(s,a) = \langle \phi(s,a),\, \widehat{w}_{h,f}^k \rangle$.

Ellipsoidal exploration bonus. The exploration bonus is

$$b_h^k(s,a) \;\coloneqq\; \beta \cdot \|\phi(s,a)\|_{(\Lambda_h^k)^{-1}}, \tag{3}$$

where $\|x\|_M \coloneqq \sqrt{x^\mathsf{T} M x}$ is the Mahalanobis norm and $\beta = \widetilde{O}(dH)$ is a confidence parameter set below.

Algorithm 2 Lin-UCBVI (episodic, linear MDP) (Jin et al., 2020)

Require: Failure probability δ, regularization $\lambda \geq 1$, confidence parameter $\beta > 0$

Initialize $\Lambda_h^0 \leftarrow \lambda I$ for all h
for $k = 0, 1, \ldots, K-1$ **do**
  Set $\widehat{V}_H^k(\cdot) \equiv 0$
  for $h = H-1, H-2, \ldots, 0$ **do** ▷ *backward induction*
    $\widehat{w}^k_{h,\widehat{V}^k_{h+1}} \leftarrow (\Lambda_h^k)^{-1} \sum_{i=0}^{k-1} \phi(s_h^i, a_h^i)\, \widehat{V}_{h+1}^k(s_{h+1}^i)$ ▷ *ridge regression*
    $\widehat{Q}_h^k(s,a) \leftarrow \max\{0, \min\{H-h,\ r_h(s,a) + \langle \widehat{w}^k_{h,\widehat{V}^k_{h+1}}, \phi(s,a)\rangle + \beta \|\phi(s,a)\|_{(\Lambda_h^k)^{-1}}\}\}$ for all (s,a)
    $\widehat{V}_h^k(s) \leftarrow \max_a \widehat{Q}_h^k(s,a), \quad \pi_h^k(s) \leftarrow \operatorname{argmax}_a \widehat{Q}_h^k(s,a)$
  end for
  Execute π^k: set $s_0^k = s_0$; for $h = 0, \ldots, H-1$, take $a_h^k = \pi_h^k(s_h^k)$, sample $s_{h+1}^k \sim P_h^\star(\cdot \mid s_h^k, a_h^k)$
  Update: $\Lambda_h^{k+1} \leftarrow \Lambda_h^k + \phi(s_h^k, a_h^k)\, \phi(s_h^k, a_h^k)^\top$ for all h
end for

Optimistic planning. Initialize $\widehat{V}_H^k(\cdot) \equiv 0$. For $h = H-1, \ldots, 0$, compute

$$\begin{aligned} \widehat{Q}_h^k(s,a) &\coloneqq \mathrm{clip}_{[0,H-h]}\Big(r_h(s,a) + (\widehat{P}_h^k \widehat{V}_{h+1}^k)(s,a) + b_h^k(s,a)\Big), \\ \widehat{V}_h^k(s) &\coloneqq \max_{a \in \mathcal{A}} \widehat{Q}_h^k(s,a). \end{aligned} \tag{4}$$

Here $(\widehat{P}_h^k \widehat{V}_{h+1}^k)(s,a) = \langle \phi(s,a),\ \widehat{w}^k_{h,\widehat{V}^k_{h+1}} \rangle$.

Optimization assumption. We assume the maximization over $a \in \mathcal{A}$ is tractable (e.g., $\mathcal{A}$ is finite or we have access to an optimization oracle).

Lin-UCBVI optimistic Q-function:

$$\widehat{Q}_h^k(s,a) = \mathrm{clip}_{[0,H-h]}\left(r_h(s,a) + \underbrace{(\widehat{P}_h^k \widehat{V}_{h+1}^k)(s,a)}_{\text{ridge regression estimate}} + \underbrace{\beta \cdot \|\phi(s,a)\|_{(\Lambda_h^k)^{-1}}}_{\text{ellipsoidal bonus}} \right).$$

From count-based to ellipsoidal bonuses. The Mahalanobis norm $\|\phi(s,a)\|_{(\Lambda_h^k)^{-1}}$ is the natural generalization of the count-based bonus $1/\sqrt{N_h^k(s,a)}$ from the tabular case. To see this, consider the tabular setting with $\phi(s,a) = e_{(s,a)} \in \mathbb{R}^{SA}$. Then $\Lambda_h^k = \lambda I + \mathrm{diag}(N_h^k(s,a))$, which is diagonal, and

$$\|\phi(s,a)\|_{(\Lambda_h^k)^{-1}} = \frac{1}{\sqrt{\lambda + N_h^k(s,a)}} \approx \frac{1}{\sqrt{N_h^k(s,a)}}$$

for $\lambda = 1$ and $N_h^k(s,a) \gg 1$. The ellipsoidal bonus is large in directions of feature space that have not been well-explored, measured by the inverse design matrix $(\Lambda_h^k)^{-1}$.

Regret Guarantee

Theorem 15 (Lin-UCBVI regret)**.** *Assume the linear transition model (Definition 2) and bounded rewards $r_h(s,a) \in [0,1]$. Run Lin-UCBVI (Algorithm 2) with regularization $\lambda \geq 1$ and bonus $b_h^k(s,a) = \beta\|\phi(s,a)\|_{(\Lambda_h^k)^{-1}}$. Fix $\delta \in (0,1)$, and suppose β is chosen so that the uniform confidence event in Proposition 16 holds with probability at least $1-\delta$. Then the expected regret satisfies*

$$R_K \;\leq\; c\,\beta H\sqrt{Kd\log\left(1+\frac{K}{\lambda}\right)} + \delta KH$$

for a universal constant $c > 0$. In particular, taking $\beta = \widetilde{O}(Hd)$ yields

$$R_K \;=\; \widetilde{O}\left(H^2\sqrt{d^3K}\right).$$

Lin-UCBVI regret:

$$R_K = \widetilde{O}\left(H^2\sqrt{d^3K}\right).$$

The bound has *no dependence* on $|\mathcal{S}|$ or $|\mathcal{A}|$, scaling only with the feature dimension d, horizon H, and episodes K.

Comparison with tabular. Setting $d = SA$ in the Lin-UCBVI bound gives $\widetilde{O}(H^2\sqrt{S^3A^3K})$, which is worse than the tabular UCBVI bound $\widetilde{O}(H^2S\sqrt{AK})$. This is the price of generality: the linear MDP framework does not exploit the special tabular structure (where the design matrix Λ is diagonal). However, for $d \ll SA$, the linear MDP bound is dramatically better—it enables polynomial regret even when $|\mathcal{S}|$ or $|\mathcal{A}|$ is exponentially large or infinite.

Analysis Overview

The proof follows the same four-step template as the tabular case: uniform confidence, optimism, a one-episode bonus bound, and summation via an elliptical potential argument.

Step 1: Uniform confidence. The key uniformization step is the following external ingredient, which we quote without proof.

Proposition 16 (Uniform model confidence)**.** *Fix $\delta \in (0,1)$. With probability at least $1-\delta$, simultaneously for all episodes $k \in \{0,\dots,K-1\}$, all stages $h \in \{0,\dots,H-1\}$, all $(s,a) \in \mathcal{S}\times\mathcal{A}$, and for $f = \widehat{V}_{h+1}^k$ produced by Lin-UCBVI,*

$$(\widehat{P}_h^k f)(s,a) - (P_h^\star f)(s,a) \;\le\; \beta \cdot \|\phi(s,a)\|_{(\Lambda_h^k)^{-1}} \;=\; b_h^k(s,a).$$

Proof idea. This is the subtle point in the original proof: the target $f = \widehat{V}_{h+1}^k$ is random and data-dependent, so a fixed-target regression bound does not suffice. Instead, Agarwal et al. (2021) apply their Corollary A.17 stage-wise to an optimistic value-function class large enough to contain all iterates of Lin-UCBVI, and then union bound over h.

Why no $\sqrt{S}$ factor. Unlike the tabular case, we never need the ℓ_1 bound. Bellman closure ensures that $P_h^\star f$ is *exactly* linear in ϕ for *any* f—including the algorithm's own random $\widehat{V}_{h+1}^k$. This is why the confidence bound applies directly to the random target, and no $\sqrt{S}$ factor arises.

Step 2: Optimism. On the event of Proposition 16, the same backward-induction argument as in the tabular case gives:

Lemma 17 (Optimism)**.** *For every episode k, every stage h, and every state s,*

$$\widehat{V}_h^k(s) \;\ge\; V_h^\star(s).$$

Proof. Fix episode k and argue by backward induction on h. At $h = H$, $\widehat{V}_H^k \equiv 0 = V_H^\star$. Assume $\widehat{V}_{h+1}^k \ge V_{h+1}^\star$ pointwise. Fix (s,a). Since $P_h^\star(\cdot \mid s,a)$ is a distribution and $\widehat{V}_{h+1}^k \ge V_{h+1}^\star$,

$$(P_h^\star V_{h+1}^\star)(s,a) \;\le\; (P_h^\star \widehat{V}_{h+1}^k)(s,a).$$

By Proposition 16 applied to $f = \widehat{V}_{h+1}^k$,

$$(P_h^\star \widehat{V}_{h+1}^k)(s,a) \;\le\; (\widehat{P}_h^k \widehat{V}_{h+1}^k)(s,a) + b_h^k(s,a).$$

Therefore,

$$Q_h^\star(s,a) = r_h(s,a) + (P_h^\star V_{h+1}^\star)(s,a) \;\le\; r_h(s,a) + (\widehat{P}_h^k \widehat{V}_{h+1}^k)(s,a) + b_h^k(s,a).$$

Since $Q_h^\star(s,a) \in [0, H-h]$, clipping to $[0, H-h]$ preserves the inequality and gives $Q_h^\star(s,a) \le \widehat{Q}_h^k(s,a)$. Taking maxima over a yields $\widehat{V}_h^k(s) \ge V_h^\star(s)$. □

Step 3: One-episode bonus bound. The per-episode regret is controlled purely by the bonuses:

Lemma 18 (Per-episode bonus bound)**.** *On the event of Proposition 16, for every episode k,*

$$V_0^\star(s_0) - V_0^{\pi^k}(s_0) \;\le\; 2\sum_{h=0}^{H-1} \mathbb{E}_{(s_h,a_h)\sim d_h^{\pi^k}}\left[b_h^k(s_h,a_h)\right].$$

Proof. Fix episode k and abbreviate $\pi = \pi^k$, $\widehat{V}_h = \widehat{V}_h^k$, $\widehat{P}_h = \widehat{P}_h^k$, and $b_h = b_h^k$. Optimism (Lemma 17) gives

$$V_0^\star(s_0) - V_0^\pi(s_0) \;\le\; \widehat{V}_0(s_0) - V_0^\pi(s_0).$$

Define $\Delta_h(s) \coloneqq \widehat{V}_h(s) - V_h^\pi(s)$, so $\Delta_H \equiv 0$.

We first show that for every stage h and state s,

$$\Delta_h(s) \;\le\; b_h(s,\pi_h(s)) \;+\; \big((\widehat{P}_h - P_h^\star)\widehat{V}_{h+1}\big)(s,\pi_h(s)) \;+\; (P_h^\star \Delta_{h+1})(s,\pi_h(s)).$$

Indeed, on the event of Proposition 16,

$$(\widehat{P}_h \widehat{V}_{h+1})(s,a) \;\ge\; (P_h^\star \widehat{V}_{h+1})(s,a) - b_h(s,a) \;\ge\; -b_h(s,a),$$

since $\widehat{V}_{h+1}(\cdot) \in [0, H-h-1]$. Recall from (4) that

$$\widehat{Q}_h(s,a) = \mathrm{clip}_{[0,H-h]}\Big(r_h(s,a) + (\widehat{P}_h \widehat{V}_{h+1})(s,a) + b_h(s,a)\Big), \qquad \widehat{V}_h(s) = \max_{a'} \widehat{Q}_h(s,a').$$

The quantity inside the clip satisfies

$$r_h(s,a) + (\widehat{P}_h \widehat{V}_{h+1})(s,a) + b_h(s,a) \;\ge\; r_h(s,a) \;\ge\; 0,$$

so clipping can only truncate from above. Therefore

$$\widehat{Q}_h(s,a) \;\le\; r_h(s,a) + (\widehat{P}_h \widehat{V}_{h+1})(s,a) + b_h(s,a).$$

Evaluating at $\pi_h(s) \in \arg\max_a \widehat{Q}_h(s,a)$,

$$\widehat{V}_h(s) = \widehat{Q}_h(s,\pi_h(s)) \;\le\; r_h(s,\pi_h(s)) + (\widehat{P}_h\widehat{V}_{h+1})(s,\pi_h(s)) + b_h(s,\pi_h(s)).$$

On the other hand,

$$V_h^\pi(s) = r_h(s,\pi_h(s)) + (P_h^\star V_{h+1}^\pi)(s,\pi_h(s)).$$

Subtracting and writing $\widehat{V}_{h+1} = V_{h+1}^\pi + \Delta_{h+1}$ gives the displayed recursion.

Taking expectations along a trajectory of π in the true MDP and unrolling from $h=0$ to $H-1$ yields

$$\widehat{V}_0(s_0) - V_0^\pi(s_0) \;\le\; \sum_{h=0}^{H-1} \mathbb{E}_{(s_h,a_h)\sim d_h^\pi}\Big[b_h(s_h,a_h) \;+\; \big((\widehat{P}_h - P_h^\star)\widehat{V}_{h+1}\big)(s_h,a_h)\Big].$$

Applying Proposition 16 once more,

$$\big((\widehat{P}_h - P_h^\star)\widehat{V}_{h+1}\big)(s,a) \;\le\; b_h(s,a),$$

so

$$\widehat{V}_0(s_0) - V_0^\pi(s_0) \;\le\; 2\sum_{h=0}^{H-1} \mathbb{E}_{(s_h,a_h)\sim d_h^\pi}[b_h(s_h,a_h)]\,.$$

Combining with the initial optimism bound gives the claim. □

Step 4: Summation via an elliptical potential. The counting lemma is replaced by:

Lemma 19 (Elliptical potential lemma)**.** *For any sequence of vectors $x_0,\dots,x_{K-1} \in \mathbb{R}^d$ with $\|x_k\| \le 1$, let $\Sigma_k = \lambda I + \sum_{i=0}^{k-1} x_i x_i^\mathsf{T}$. Then*

$$\sum_{k=0}^{K-1} \|x_k\|^2_{\Sigma_k^{-1}} \;\le\; 2d\log\left(1 + \frac{K}{\lambda}\right).$$

Proof. Since $\lambda \ge 1$ and $\|x_k\| \le 1$, we have $\|x_k\|^2_{\Sigma_k^{-1}} \le 1$ for all k. By the matrix determinant lemma,

$$\det(\Sigma_{k+1}) = \det(\Sigma_k)\big(1 + x_k^\mathsf{T}\Sigma_k^{-1}x_k\big),$$

so

$$\det(\Sigma_K) = \det(\Sigma_0)\prod_{k=0}^{K-1}\big(1 + \|x_k\|^2_{\Sigma_k^{-1}}\big).$$

Using $\|x_k\|^2_{\Sigma_k^{-1}} \le 2\log(1+\|x_k\|^2_{\Sigma_k^{-1}})$ on $[0,1]$,

$$\sum_{k=0}^{K-1} \|x_k\|^2_{\Sigma_k^{-1}} \;\le\; 2\sum_{k=0}^{K-1} \log\bigl(1+\|x_k\|^2_{\Sigma_k^{-1}}\bigr) \;=\; 2\log\frac{\det(\Sigma_K)}{\det(\Sigma_0)}.$$

Finally, $\Sigma_K = \lambda I + \sum_{k=0}^{K-1} x_k x_k^\mathsf{T} \preceq (\lambda+K)I$, hence

$$\frac{\det(\Sigma_K)}{\det(\Sigma_0)} \;\le\; \left(1+\frac{K}{\lambda}\right)^d,$$

which gives the claim. □

Elliptical potential lemma (linear analog of the counting lemma):

$$\sum_{k=0}^{K-1} \|\phi(s_h^k,a_h^k)\|^2_{(\Lambda_h^k)^{-1}} \;\le\; 2d\log\left(1+\frac{K}{\lambda}\right) \qquad \text{for each fixed stage } h.$$

Completing the proof of Theorem 15. Let E denote the event in Proposition 16. On E, Lemma 18 implies that for each episode k,

$$V_0^\star(s_0) - V_0^{\pi^k}(s_0) \;\le\; 2\sum_{h=0}^{H-1} \mathbb{E}\bigl[b_h^k(s_h^k,a_h^k)\,\big|\,\mathcal{H}_{<k}\bigr],$$

since b_h^k is $\mathcal{H}_{<k}$-measurable and (s_h^k,a_h^k) has occupancy distribution under π^k in the true MDP. Therefore, using $\mathbb{1}\{E\} \le 1$ and the tower property,

$$\mathbb{E}\left[\mathbb{1}\{E\}\sum_{k=0}^{K-1}\bigl(V_0^\star(s_0) - V_0^{\pi^k}(s_0)\bigr)\right] \;\le\; 2\mathbb{E}\left[\sum_{k=0}^{K-1}\sum_{h=0}^{H-1} b_h^k(s_h^k,a_h^k)\right].$$

Substituting $b_h^k(s,a) = \beta\|\phi(s,a)\|_{(\Lambda_h^k)^{-1}}$ gives

$$\mathbb{E}\left[\mathbb{1}\{E\}\sum_{k=0}^{K-1}\bigl(V_0^\star(s_0) - V_0^{\pi^k}(s_0)\bigr)\right] \;\le\; 2\beta\sum_{h=0}^{H-1}\mathbb{E}\left[\sum_{k=0}^{K-1}\|\phi(s_h^k,a_h^k)\|_{(\Lambda_h^k)^{-1}}\right].$$

Fix a stage h and write $x_k \coloneqq \phi(s_h^k,a_h^k)$ and $\Sigma_k \coloneqq \Lambda_h^k$. By Cauchy–Schwarz and Lemma 19,

$$\sum_{k=0}^{K-1}\|x_k\|_{\Sigma_k^{-1}} \;\le\; \sqrt{K\sum_{k=0}^{K-1}\|x_k\|^2_{\Sigma_k^{-1}}} \;\le\; \sqrt{2Kd\log\left(1+\frac{K}{\lambda}\right)}.$$

Summing over $h = 0, \dots, H-1$ yields

$$\mathbb{E}\left[\mathbb{1}\{E\} \sum_{k=0}^{K-1} \left(V_0^\star(s_0) - V_0^{\pi^k}(s_0)\right)\right] \;\le\; c\,\beta H \sqrt{Kd \log\left(1 + \frac{K}{\lambda}\right)}$$

for a universal constant $c > 0$. On $\overline{E}$, the trivial bound $0 \le V_0^\star(s_0) - V_0^{\pi^k}(s_0) \le H$ contributes at most $\delta K H$. Combining proves the theorem.

Beyond Pointwise Optimism: The GOLF Algorithm

Notation change. To align with the original references (Jin et al., 2021; Xie et al., 2023), the remainder of this lecture adopts a slightly different convention: steps are indexed $h = 1, \dots, H$ (instead of $0, \dots, H-1$), the initial state is s_1 (instead of s_0), and rewards are normalized so that $\sum_{h=1}^{H} r_h(s_h, a_h) \in [0, 1]$ for every trajectory. Under this normalization, $Q_h^\star : \mathcal{S} \times \mathcal{A} \to [0, 1]$ for all h, and the Bellman operator is $(\mathcal{T}_h g)(s, a) = r_h(s, a) + \mathbb{E}_{s' \sim P_h^\star(\cdot|s,a)}[\max_{a'} g(s', a')]$.

Motivation: Why Pointwise Bonuses Break Down

Both UCBVI and Lin-UCBVI achieve optimism by adding an exploration bonus to each state-action pair:

$$\widehat{Q}_h^k(s, a) = r_h(s, a) + \widehat{P}_h^k \widehat{V}_{h+1}^k(s, a) + b_h^k(s, a).$$

The bonus is designed so that $\widehat{Q}_h^k(s, a) \ge Q_h^\star(s, a)$ for *all* (s, a) simultaneously—what we call *pointwise optimism*. For tabular MDPs, the bonus $b = O(H/\sqrt{N})$ is a simple function of the visit count. For linear MDPs, the bonus $b = \beta \|\phi\|_{(\Lambda^k)^{-1}}$ is a closed-form function of the feature vector and the design matrix.

But what about general function approximation? Suppose we are given a function class $\mathcal{F}$ of candidate Q-functions (e.g., neural networks, decision trees, or any other parametric family). There is no natural way to construct a pointwise bonus $b_h^k(s, a)$ such that $f_h(s, a) + b_h^k(s, a) \ge Q_h^\star(s, a)$ for all (s, a), while also ensuring that $f_h + b_h^k$ remains within a tractable function class. The bonus depends on the model structure in a way that does not generalize.

Pointwise vs. global optimism.

	Pointwise optimism (UCBVI, Lin-UCBVI)	**Global optimism** (GOLF)
Goal	$\widehat{Q}_h^k(s,a) \geq Q_h^\star(s,a)$ for *all* (s,a,h)	$\widehat{V}_1^k(s_1) \geq V_1^\star(s_1)$ at the *initial state only*
Mechanism	Add explicit bonus $b_h^k(s,a)$ to each state-action	Search for most optimistic value function in a confidence set $\mathcal{B}^k$
Proof of optimism	Backward induction on h	Confidence set containment: $Q^\star \in \mathcal{B}^k$
Scope	Requires closed-form bonus (tabular, linear)	Works for any function class $\mathcal{F}$

Global optimism is strictly weaker than pointwise optimism: it only guarantees optimism at s_1, not everywhere. Yet this is sufficient for bounding regret, since regret only involves $V_1^\star(s_1) - V_1^{\pi^k}(s_1)$.

Setup

In this section, for simplicity, we assume $|\mathcal{F}| < \infty$ and define the confidence set using the same class $\mathcal{F}_h$ on both sides of the local least-squares comparison. This is the common setting used below to present both the BE-dimension analysis of Jin et al. (2021) and the coverability analysis of Xie et al. (2023).

We are given a value function class $\mathcal{F} = \mathcal{F}_1 \times \cdots \times \mathcal{F}_H$, where each $\mathcal{F}_h \subseteq (\mathcal{S} \times \mathcal{A} \to [0,1])$ is finite. We adopt the convention $f_{H+1} \equiv 0$. For each $f \in \mathcal{F}$, let π_f denote its greedy policy:

$$\pi_{f,h}(s) \in \operatorname*{argmax}_{a \in \mathcal{A}} f_h(s,a).$$

We write $\Pi \coloneqq \{\pi_f : f \in \mathcal{F}\}$ for the induced policy class.

Assumption 1 (Completeness). *For all $h \in [H]$ and all $f_{h+1} \in \mathcal{F}_{h+1}$, we have*

$$\mathcal{T}_h f_{h+1} \in \mathcal{F}_h.$$

Algorithm: GOLF

GOLF (**G**lobal **O**ptimism based on **L**ocal **F**itting) replaces explicit pointwise bonuses with a *confidence set* of Q-functions.

Algorithm 3 GOLF

Require: Function class $\mathcal{F}$, confidence parameter β, episodes K
Initialize $\mathcal{D}_h \leftarrow \emptyset$ for all $h \in [H]$, and $\mathcal{B}^0 \leftarrow \mathcal{F}$
for $k = 1, 2, \ldots, K$ **do**
  $f^k \leftarrow \operatorname{argmax}_{f \in \mathcal{B}^{k-1}} \max_a f_1(s_1, a)$
  Set $\pi^k \leftarrow \pi_{f^k}$
  Execute π^k, collect trajectory $(s_1^k, a_1^k, r_1^k, \ldots, s_H^k, a_H^k, r_H^k, s_{H+1}^k)$
  $\mathcal{D}_h \leftarrow \mathcal{D}_h \cup \{(s_h^k, a_h^k, r_h^k, s_{h+1}^k)\}$ for all $h \in [H]$
  Update $\mathcal{B}^k = \{f \in \mathcal{F} : \mathcal{L}_{\mathcal{D}_h}(f_h, f_{h+1}) \le \min_{g_h \in \mathcal{F}_h} \mathcal{L}_{\mathcal{D}_h}(g_h, f_{h+1}) + \beta,\ \forall h \in [H]\}$
end for
return π^{out} sampled uniformly from $\{\pi^1, \ldots, \pi^K\}$

Squared Bellman error loss. For each step h, let $\mathcal{D}_h = \{(s_h^i, a_h^i, r_h^i, s_{h+1}^i)\}_{i=1}^{k-1}$ be the data collected at step h from the first $k-1$ episodes. For $f_h \in \mathcal{F}_h$ and $f_{h+1} \in \mathcal{F}_{h+1}$, define

$$\mathcal{L}_{\mathcal{D}_h}(f_h, f_{h+1}) \coloneqq \sum_{(s,a,r,s') \in \mathcal{D}_h} \left[f_h(s,a) - r - \max_{a' \in \mathcal{A}} f_{h+1}(s', a') \right]^2.$$

Confidence set. At the start of episode k, GOLF keeps all functions whose Bellman loss is within β of the best step-wise fit:

GOLF **confidence set:**

$$\mathcal{B}^k = \left\{ f \in \mathcal{F} : \mathcal{L}_{\mathcal{D}_h}(f_h, f_{h+1}) \le \min_{g_h \in \mathcal{F}_h} \mathcal{L}_{\mathcal{D}_h}(g_h, f_{h+1}) + \beta, \quad \forall h \in [H] \right\}.$$

Optimistic planning. GOLF selects the most optimistic value function in $\mathcal{B}^{k-1}$ and executes its greedy policy:

$$f^k \in \operatorname*{argmax}_{f \in \mathcal{B}^{k-1}} \max_{a \in \mathcal{A}} f_1(s_1, a), \qquad \pi^k = \pi_{f^k}.$$

Interpretation. GOLF is an optimistic version of fitted Q-iteration: instead of taking a single least-squares fit, it keeps the whole version space of functions whose Bellman loss is nearly optimal, and then chooses the most optimistic element.

Shared High-Probability Ingredients

For each episode k and step h, define the Bellman residual

$$\delta_h^k(s,a) \coloneqq f_h^k(s,a) - (\mathcal{T}_h f_{h+1}^k)(s,a),$$

the average Bellman error under policy π,

$$\mathcal{E}(f^k, \pi, h) \;:=\; \mathbb{E}_{(s,a)\sim d_h^\pi}\left[\delta_h^k(s,a)\right],$$

and the cumulative visitation

$$\widetilde{d}_h^{(k)}(s,a) \;:=\; \sum_{i=1}^{k-1} d_h^{\pi^i}(s,a).$$

We quote the following high-probability event without proof. In our setting it is the standard concentration event behind both analyses: compare Lemmas 39–40 and Appendix D.3 in Jin et al. (2021), and Eq. (2) in Section 3.2 of Xie et al. (2023).

Lemma 20 (Confidence Set Containment)**.** *There is an absolute constant c such that if*

$$\beta = c\log(KH|\mathcal{F}|/\delta),$$

then with probability at least $1-\delta$, we have

$$Q^\star \in \mathcal{B}^k \qquad \textit{for all } k \in [K].$$

Lemma 21 (Bounded Squared Bellman Error)**.** *Under the same choice of β, with probability at least $1-\delta$, for all $(k,h) \in [K]\times[H]$:*

(a) $\displaystyle\sum_{i=1}^{k-1} \mathbb{E}_{(s,a)\sim d_h^{\pi^i}}\left[(\delta_h^k(s,a))^2\right] \le O(\beta),$

(b) $\displaystyle\sum_{i=1}^{k-1} (\delta_h^k(s_h^i, a_h^i))^2 \le O(\beta).$

Equivalently,

$$\sum_{s,a} \widetilde{d}_h^{(k)}(s,a)\,(\delta_h^k(s,a))^2 \le O(\beta).$$

Lemma 22 (Regret Decomposition Into Bellman Errors)**.** *On the event of Lemma 20,*

$$R_K \;\le\; \sum_{k=1}^{K}\sum_{h=1}^{H} \mathcal{E}(f^k, \pi^k, h).$$

Proof. Since $Q^\star \in \mathcal{B}^{k-1}$ and f^k is chosen from $\mathcal{B}^{k-1}$, optimistic planning gives

$$\max_a f_1^k(s_1,a) \;\ge\; \max_a Q_1^\star(s_1,a) \;=\; V_1^\star(s_1).$$

By the standard policy loss decomposition,

$$\max_a f_1^k(s_1,a) - V_1^{\pi^k}(s_1) \;=\; \sum_{h=1}^{H} \mathbb{E}_{(s,a)\sim d_h^{\pi^k}}\left[f_h^k(s,a) - (\mathcal{T}_h f_{h+1}^k)(s,a)\right] \;=\; \sum_{h=1}^{H} \mathcal{E}(f^k,\pi^k,h).$$

Hence

$$V_1^\star(s_1) - V_1^{\pi^k}(s_1) \;\le\; \sum_{h=1}^{H} \mathcal{E}(f^k,\pi^k,h),$$

and summing over k proves the claim. □

Shared reduction:

$$R_K \;\le\; \sum_{k=1}^{K}\sum_{h=1}^{H} \mathcal{E}(f^k,\pi^k,h).$$

Regret via BE Dimension

This subsection follows the BE-dimension analysis of Jin et al. (2021, Section 4.2 and Appendix D).

Definition 3 (Distributional Eluder Dimension). *Let Φ be a function class on $\mathcal{X}$, and Π a family of probability measures over $\mathcal{X}$. We say ν is ε-independent of $\{\mu_1,\dots,\mu_n\} \subset \Pi$ with respect to Φ if there exists $\phi \in \Phi$ such that*

$$\sqrt{\sum_{i=1}^{n}(\mathbb{E}_{\mu_i}[\phi])^2} \le \varepsilon \qquad \textit{but} \qquad |\mathbb{E}_\nu[\phi]| > \varepsilon.$$

The distributional Eluder dimension $\dim_{\mathrm{DE}}(\Phi,\Pi,\varepsilon)$ is the length of the longest sequence in Π that is successively ε'-independent for some $\varepsilon' \ge \varepsilon$.

Definition 4 (Bellman Eluder Dimension). *For a distribution family $\Pi = \{\Pi_h\}_{h\in[H]}$, define*

$$d_{\mathrm{BE}}(\mathcal{F},\Pi,\varepsilon) \;\coloneqq\; \max_{h\in[H]} \dim_{\mathrm{DE}}((I-\mathcal{T}_h)\mathcal{F},\Pi_h,\varepsilon)\,,$$

where

$$(I-\mathcal{T}_h)\mathcal{F} \;=\; \{f_h - \mathcal{T}_h f_{h+1} : f \in \mathcal{F}\}.$$

We use the two distribution families that appear in the BE-dimension analysis:

$$\mathcal{D}_{\mathcal{F}} = \{\mathcal{D}_{\mathcal{F},h}\}_{h=1}^{H}, \qquad \mathcal{D}_{\mathcal{F},h} \coloneqq \{d_h^{\pi_f} : f \in \mathcal{F}\},$$

and

$$\mathcal{D}_\Delta = \{\mathcal{D}_{\Delta,h}\}_{h=1}^H, \qquad \mathcal{D}_{\Delta,h} \coloneqq \{\delta_{(s,a)}(\cdot) : (s,a) \in \mathcal{S} \times \mathcal{A}\}.$$

Lemma 23 (DE-Dimension Conversion). *Let Φ be a function class on $\mathcal{X}$ with $|\phi(x)| \le C$ for all $(\phi, x) \in \Phi \times \mathcal{X}$, and let Π be a family of probability measures over $\mathcal{X}$. Suppose sequences $\{\phi_k\}_{k=1}^K \subset \Phi$ and $\{\mu_k\}_{k=1}^K \subset \Pi$ satisfy*

$$\sum_{i=1}^{k-1} (\mathbb{E}_{\mu_i}[\phi_k])^2 \le \beta \qquad \text{for all } k \in [K].$$

Then for all $\omega > 0$ and all $k \in [K]$,

$$\sum_{i=1}^{k} |\mathbb{E}_{\mu_i}[\phi_i]| \;\le\; O\Big(\sqrt{\dim_{\mathrm{DE}}(\Phi, \Pi, \omega)\, \beta\, k} \;+\; \min\{k, \dim_{\mathrm{DE}}(\Phi, \Pi, \omega)\}\, C \;+\; k\omega\Big).$$

Theorem 24 (GOLF Regret via BE Dimension). *Under Assumption 1, run* GOLF *(Algorithm 3) with*

$$\beta = c \log(KH|\mathcal{F}|/\delta)$$

for a sufficiently large absolute constant c. Then with probability at least $1-\delta$,

$$R_K \;\le\; O\Big(H\sqrt{d_{\mathrm{BE}}\, \beta\, K}\Big), \qquad d_{\mathrm{BE}} \coloneqq \min_{\Pi \in \{\mathcal{D}_\mathcal{F}, \mathcal{D}_\Delta\}} d_{\mathrm{BE}}(\mathcal{F}, \Pi, 1/\sqrt{K}).$$

Proof. On the event of Lemmas 20–21,

$$R_K \;\le\; \sum_{k=1}^K \sum_{h=1}^H \mathcal{E}(f^k, \pi^k, h) \;\le\; \sum_{h=1}^H \sum_{k=1}^K |\mathcal{E}(f^k, \pi^k, h)|.$$

Fix a step h. We prove the branch corresponding to $\Pi = \mathcal{D}_\mathcal{F}$; the $\mathcal{D}_\Delta$ branch is analogous. Apply Lemma 23 with

$$\mathcal{X} = \mathcal{S} \times \mathcal{A}, \qquad \Phi = (I - \mathcal{T}_h)\mathcal{F}, \qquad \phi_k = \delta_h^k, \qquad \mu_k = d_h^{\pi^k}, \qquad C = 1, \qquad \omega = 1/\sqrt{K}.$$

By Jensen's inequality and Lemma 21(a),

$$\sum_{i=1}^{k-1} (\mathbb{E}_{\mu_i}[\phi_k])^2 = \sum_{i=1}^{k-1} \big(\mathcal{E}(f^k, \pi^i, h)\big)^2 \le \sum_{i=1}^{k-1} \mathbb{E}_{(s,a)\sim d_h^{\pi^i}}\big[(\delta_h^k(s,a))^2\big] \le O(\beta).$$

Therefore

$$\sum_{k=1}^K |\mathcal{E}(f^k, \pi^k, h)| \;\le\; O\Big(\sqrt{d_{\mathrm{BE}}\, \beta\, K} + d_{\mathrm{BE}} + \sqrt{K}\Big).$$

Summing over $h = 1, \dots, H$ and absorbing lower-order terms proves the theorem. □

GOLF **regret (BE dimension):**

$$R_K = \widetilde{O}\left(H\sqrt{d_{\mathrm{BE}}\, K}\right).$$

Regret via Coverability

This subsection follows the coverability analysis of Xie et al. (2023, Section 3.2) for the same GOLF algorithm.

Definition 5 (Coverability Coefficient)**.** *The coverability coefficient of the policy class* Π *is*

$$C_{\mathrm{cov}} \coloneqq \inf_{\mu_1,\dots,\mu_H \in \Delta(\mathcal{S}\times\mathcal{A})} \sup_{\pi\in\Pi,\, h\in[H]} \left\| \frac{d_h^\pi}{\mu_h} \right\|_\infty .$$

Lemma 25 (Coverability as Cumulative Reachability)**.** *The coverability coefficient also admits the equivalent form*

$$C_{\mathrm{cov}} = \max_{h\in[H]} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \sup_{\pi\in\Pi} d_h^\pi(s,a).$$

Burn-in times. For each step h, let $\mu_h^\star$ be a distribution attaining the infimum in Definition 5; if the infimum is not attained, one may use an arbitrarily close minimizer and absorb constants. For each state-action pair (s,a), define

$$\tau_h(s,a) \coloneqq \min\left(\left\{k\in[K] : \widetilde{d}_h^{(k)}(s,a) \geq C_{\mathrm{cov}}\, \mu_h^\star(s,a)\right\} \cup \{K+1\}\right). \tag{5}$$

Lemma 26 (Per-State-Action Elliptic Potential)**.** *Let* $d^{(1)}, \dots, d^{(T)}$ *be distributions over a set* $\mathcal{Z}$*, and let* $\mu \in \Delta(\mathcal{Z})$ *satisfy*

$$\frac{d^{(t)}(z)}{\mu(z)} \leq C \qquad \textit{for all } (z,t) \in \mathcal{Z}\times[T].$$

Then for every $z \in \mathcal{Z}$*,*

$$\sum_{t=1}^{T} \frac{d^{(t)}(z)}{\sum_{i<t} d^{(i)}(z) + C\,\mu(z)} \leq O(\log T).$$

Proof. Fix $z \in \mathcal{Z}$, and write $w_t = d^{(t)}(z)$, $W_t = \sum_{i<t} w_i$, and $W = C\mu(z)$. Since $w_t \leq W$, we have $w_t/(W_t+W) \in [0,1]$. Using $u \leq 2\log(1+u)$ for $u \in [0,1]$,

$$\frac{w_t}{W_t+W} \leq 2\log\left(1+\frac{w_t}{W_t+W}\right) = 2\log\left(\frac{W_{t+1}+W}{W_t+W}\right).$$

Summing over t telescopes:

$$\sum_{t=1}^{T} \frac{w_t}{W_t + W} \;\le\; 2\log\left(\frac{W_{T+1} + W}{W}\right) \;=\; 2\log\left(1 + \frac{\sum_{t=1}^{T} w_t}{W}\right) \;\le\; 2\log(1+T).$$

□

Theorem 27 (GOLF Regret via Coverability). *Under Assumption 1, run* GOLF *(Algorithm 3) with*

$$\beta = c\log(KH|\mathcal{F}|/\delta)$$

for a sufficiently large absolute constant c. Then with probability at least $1-\delta$,

$$R_K \;\le\; O\left(H\sqrt{C_{\mathrm{cov}}\,\beta\,K\,\log K}\right).$$

Proof. On the event of Lemmas 20–21,

$$R_K \;\le\; \sum_{k=1}^{K}\sum_{h=1}^{H} \mathcal{E}(f^k, \pi^k, h).$$

Fix a step h. Using $|\mathbb{E}[X]| \le \mathbb{E}[|X|]$, we bound

$$\sum_{k=1}^{K} \mathcal{E}(f^k, \pi^k, h)$$

by the sum of a burn-in term and a stable term (recall $\tau_h(s,a)$ from (5)):

$$\sum_{k=1}^{K} \mathbb{E}_{(s,a)\sim d_h^{\pi^k}}\left[|\delta_h^k(s,a)|\,\mathbb{1}\{k < \tau_h(s,a)\}\right] \;+\; \sum_{k=1}^{K} \mathbb{E}_{(s,a)\sim d_h^{\pi^k}}\left[|\delta_h^k(s,a)|\,\mathbb{1}\{k \ge \tau_h(s,a)\}\right].$$

Burn-in. Since $|\delta_h^k(s,a)| \le 1$,

$$\begin{aligned}\sum_{k=1}^{K} \mathbb{E}_{(s,a)\sim d_h^{\pi^k}}\left[|\delta_h^k(s,a)|\,\mathbb{1}\{k < \tau_h(s,a)\}\right] &\le \sum_{s,a}\sum_{k<\tau_h(s,a)} d_h^{\pi^k}(s,a) \\ &= \sum_{s,a} \widetilde{d}_h^{(\tau_h(s,a))}(s,a).\end{aligned}$$

By the definition of $\tau_h(s,a)$,

$$\widetilde{d}_h^{(\tau_h(s,a))}(s,a) \le 2C_{\mathrm{cov}}\,\mu_h^\star(s,a),$$

so

$$\text{Burn-in} \;\le\; \sum_{s,a} 2C_{\text{cov}}\,\mu_h^\star(s,a) \;=\; 2C_{\text{cov}}.$$

Stable phase. For $k \ge \tau_h(s,a)$ we have

$$\widetilde{d}_h^{(k)}(s,a) \ge C_{\text{cov}}\,\mu_h^\star(s,a),$$

hence

$$\widetilde{d}_h^{(k)}(s,a) \ge \frac{1}{2}\big(\widetilde{d}_h^{(k)}(s,a) + C_{\text{cov}}\,\mu_h^\star(s,a)\big).$$

Applying Cauchy–Schwarz,

$$\begin{aligned}
&\sum_{k=1}^{K} \mathbb{E}_{(s,a)\sim d_h^{\pi^k}}\left[|\delta_h^k(s,a)|\,\mathbb{1}\{k \ge \tau_h(s,a)\}\right] \\
&\le \underbrace{\sqrt{\sum_{k,s,a} \frac{\left(d_h^{\pi^k}(s,a)\,\mathbb{1}\{k \ge \tau_h(s,a)\}\right)^2}{\widetilde{d}_h^{(k)}(s,a)}}}_{\textbf{(I)}} \cdot \underbrace{\sqrt{\sum_{k,s,a} \widetilde{d}_h^{(k)}(s,a)\,(\delta_h^k(s,a))^2}}_{\textbf{(II)}}.
\end{aligned}$$

By Lemma 21,

$$\textbf{(II)}^2 \le O(\beta K), \qquad \textbf{(II)} \le O(\sqrt{\beta K}).$$

For **(I)**, on the stable phase ($k \ge \tau_h(s,a)$) we may replace $\widetilde{d}_h^{(k)}(s,a)$ in the denominator by $\frac{1}{2}(\widetilde{d}_h^{(k)}(s,a) + C_{\text{cov}}\,\mu_h^\star(s,a))$ and then drop the indicator (adding only non-negative terms):

$$\textbf{(I)}^2 \le 2\sum_{k,s,a} \frac{d_h^{\pi^k}(s,a)^2}{\widetilde{d}_h^{(k)}(s,a) + C_{\text{cov}}\,\mu_h^\star(s,a)}$$

Bounding one copy of $d_h^{\pi^k}(s,a)$ in the numerator by $\max_{k'} d_h^{\pi^{k'}}(s,a)$, then applying $\sum_{s,a} a_{s,a}\,b_{s,a} \le (\max_{s,a} a_{s,a}) \sum_{s,a} b_{s,a}$ (Hölder's inequality):

$$\le 2\left(\max_{(s,a)} \sum_{k=1}^{K} \frac{d_h^{\pi^k}(s,a)}{\widetilde{d}_h^{(k)}(s,a) + C_{\text{cov}}\,\mu_h^\star(s,a)}\right)\left(\sum_{s,a} \max_k d_h^{\pi^k}(s,a)\right).$$

For the second factor, $\max_k d_h^{\pi^k}(s,a) \le \sup_{\pi\in\Pi} d_h^{\pi}(s,a)$ pointwise, so Lemma 25 gives

$$\sum_{s,a} \max_k d_h^{\pi^k}(s,a) \le C_{\text{cov}}.$$

Lemma 26 bounds the first factor by $O(\log K)$. Therefore

$$\textbf{(I)}^2 \le O(C_{\mathrm{cov}} \log K), \qquad \textbf{(I)} \le O(\sqrt{C_{\mathrm{cov}} \log K}).$$

Combining burn-in and stable phases,

$$\sum_{k=1}^{K} \mathcal{E}(f^k, \pi^k, h) \le 2C_{\mathrm{cov}} + O\Big(\sqrt{C_{\mathrm{cov}} \log K} \cdot \sqrt{\beta K}\Big) = O\Big(\sqrt{C_{\mathrm{cov}}\, \beta\, K\, \log K}\Big).$$

Summing over $h = 1, \dots, H$ proves the theorem. □

GOLF **regret (coverability):**

$$R_K = \widetilde{O}\Big(H\sqrt{C_{\mathrm{cov}}\, K}\Big).$$

Two complementary views of the same GOLF algorithm.

	BE dimension analysis	**Coverability analysis**
Measures	Function-class complexity	MDP structural complexity
Bound	$\widetilde{O}(H\sqrt{d_{\mathrm{BE}}\, K})$	$\widetilde{O}(H\sqrt{C_{\mathrm{cov}}\, K})$
Shared inputs	same confidence set, the same optimism lemma, and the same squared Bellman-error control event	
Step 3 tool	DE-dimension conversion	Burn-in/stable decomposition + elliptic potential

The two bounds are incomparable in general: BE dimension is small when the function class is simple, while coverability is small when the family of reachable state-action distributions has strong overlap.

Summary

Summary of regret bounds:

Algorithm / Bound	Regret	Key Technique
Random exploration	$\Omega(A^H)$	Exponentially slow
UCBVI (Hoeffding)	$\widetilde{O}(H^2S\sqrt{AK})$	Pointwise bonus + counting lemma
UCBVI (Bernstein)	$\widetilde{O}(H^2\sqrt{SAK})$	Variance-dependent bonus
Lin-UCBVI (linear MDP)	$\widetilde{O}(H^2\sqrt{d^3K})$	Ellipsoidal bonus + elliptical potential
GOLF (BE dimension)	$\widetilde{O}(H\sqrt{d_{\mathrm{BE}} \cdot K})$	Confidence set + DE pigeonhole
GOLF (coverability)	$\widetilde{O}(H\sqrt{C_{\mathrm{cov}} \cdot K})$	Burn-in/stable + elliptic potential
Minimax lower bound	$\Omega(H^{3/2}\sqrt{SAK})$	Information-theoretic

Key takeaways.

- The exploration problem is fundamentally different from policy optimization: the agent must adaptively collect data online, and the "optimism in the face of uncertainty" principle guides exploration toward informative states.
- UCBVI achieves near-optimal regret by adding exploration bonuses to the empirical Bellman equation. The canonical four-step proof structure—concentration, optimism, decomposition, counting—applies broadly to optimistic exploration algorithms.
- The linear transition model extends the framework to function approximation, with the feature dimension d replacing SA. The Mahalanobis-norm bonus $\|\phi\|_{(\Lambda^k)^{-1}}$ and the elliptical potential lemma are the natural linear-algebraic analogs of count-based bonuses and the pigeonhole counting argument.
- GOLF extends the optimism principle beyond pointwise bonuses to general function classes via *global optimism*: maintaining a confidence set and selecting the most optimistic value function. Two complementary analyses yield regret bounds: the BE dimension measures function class complexity, while coverability measures MDP structural complexity. The same algorithm achieves low regret under either condition—only the analysis differs.
- The transition from generative model / offline settings (earlier lectures) to online episodic learning requires fundamentally new ideas: *optimism* replaces certainty-equivalence, and *adaptive data collection* creates coupling between the model estimates and the trajectory distribution.

References

Alekh Agarwal, Nan Jiang, Sham M Kakade, and Wen Sun. *Reinforcement Learning: Theory and Algorithms*. Self-published, 2021. Available at https://rltheorybook.github.io/.

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2–3):235–256, 2002.

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, pages 263–272, 2017.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pages 2137–2143, 2020.

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, 2021.

Tengyang Xie, Dylan J Foster, Yu Bai, Nan Jiang, and Sham M Kakade. The role of coverage in online reinforcement learning. In *International Conference on Learning Representations*, 2023.