CSE 542: Statistical Reinforcement Learning

Lecture 4: Online Reinforcement Learning

Kevin Jamieson

Paul G. Allen School of Computer Science & Engineering
University of Washington

Outline

1. The Online RL Setting
2. PAC Framework & Hardness
3. Regret & Online-to-Batch
4. Multi-Armed Bandits
5. Contextual Bandits
6. UCB Value Iteration

Today's Plan

1. The Online RL Setting
2. PAC Framework & Hardness
3. Regret & Online-to-Batch
4. Multi-Armed Bandits
5. Contextual Bandits
6. UCB Value Iteration

The Online RL Protocol

Learner interacts with a **fixed unknown MDP** $(\mathcal{S}, \mathcal{A}, \{P_h\}_h, \{r_h\}_h, H, \nu)$ for K episodes.

Episode $k = 1, 2, \ldots, K$

1. Learner selects policy $\pi^k = \{\pi_h^k\}_{h=1}^H$.
2. Environment generates trajectory:

$$s_1^k \sim \nu, \quad a_h^k = \pi_h^k(s_h^k), \quad s_{h+1}^k \sim P_h(\cdot \mid s_h^k, a_h^k).$$

3. Learner observes $(s_1^k, a_1^k, r_1^k, \ldots, s_H^k, a_H^k, r_H^k)$.

Goal: Use past experience to choose better policies — balance exploration and exploitation.

Contrast with offline RL: No fixed dataset; data is collected adaptively.

Two Performance Metrics

There are two standard ways to measure success after K episodes.

PAC Framework

Output a single policy $\widehat{\pi}$ at the end.

Goal: with prob. $\geq 1 - \delta$,

$$V_0^\star - V_0^{\widehat{\pi}} \leq \varepsilon.$$

Metric: min K to guarantee this.

Regret

Measure cumulative sub-optimality:

$$R(K) \triangleq \sum_{k=1}^{K} \left(V_0^\star - V_0^{\pi^k}\right).$$

Goal: $R(K) = \widetilde{O}(\sqrt{K})$.

Key relationship

Low regret $\Rightarrow$ good PAC guarantee (online-to-batch conversion).

Today's Plan

1. The Online RL Setting
2. PAC Framework & Hardness
3. Regret & Online-to-Batch
4. Multi-Armed Bandits
5. Contextual Bandits
6. UCB Value Iteration

The PAC Framework

PAC Definition

An algorithm is (ε, δ)-PAC with sample complexity K if, after K episodes, it outputs a policy $\widehat{\pi}$ satisfying

$$\mathbb{P}\left(V_0^\star - V_0^{\widehat{\pi}} \leq \varepsilon\right) \geq 1 - \delta.$$

What PAC cares about: quality of the *final* policy only.
What PAC ignores: cost of exploration during training — playing bad policies is fine.

Question

In a multi-armed bandit with N arms, *uniform exploration* suffices: try each arm $O(1/\varepsilon^2)$ times, pick the best. Total cost: $O(N/\varepsilon^2)$.

Why can't we do the same in an MDP?

Why Is This Hard? The Combination Lock

Combination Lock MDP

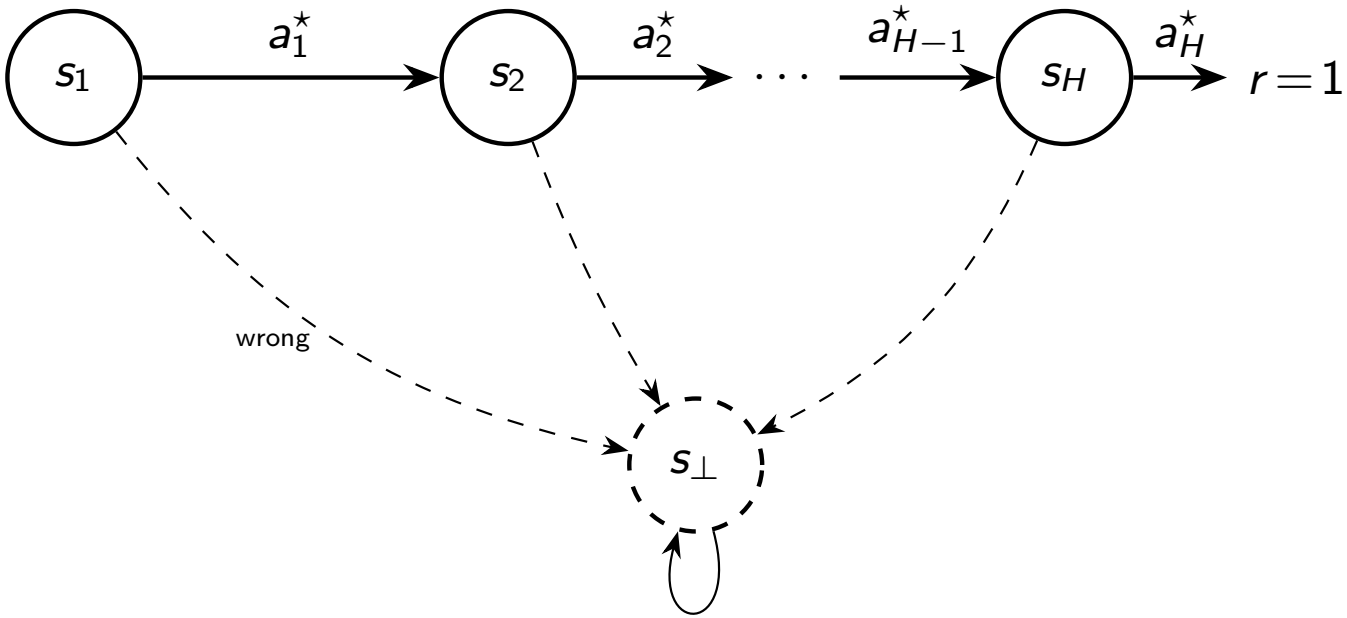
- States: $\mathcal{S} = \{s_1, \ldots, s_H, s_\perp\}$ ($s_\perp$ = absorbing dead state)
- Actions: $|\mathcal{A}| = A$; at each stage h there is one correct action $a_h^\star$
- Transitions from live state s_h:

$$P_h(\cdot \mid s_h, a) = \begin{cases} s_{h+1} & \text{if } a = a_h^\star \\ s_\perp & \text{if } a \neq a_h^\star \end{cases} \qquad P_h(s_\perp \mid s_\perp, a) = 1$$

- Rewards: all zero except $r_H(s_H, a_H^\star) = 1$
- Initial state: always s_1

Optimal value: $V_0^\star = 1$. Any policy that doesn't know the combination earns ≈ 0.

Combination Lock: Diagram & Hardness



Naive exploration fails exponentially

Uniform random policy reaches s_H with probability $(1/A)^{H-1}$ and earns reward with prob. $(1/A)^H$.

$\Rightarrow$ need $\Omega(A^H)$ episodes to see even *one* rewarding episode. **Exponential in H!**

Today's Plan

1. The Online RL Setting
2. PAC Framework & Hardness
3. Regret & Online-to-Batch
4. Multi-Armed Bandits
5. Contextual Bandits
6. UCB Value Iteration

Regret

Definition: Regret

The *regret* after K episodes is

$$R(K) \triangleq \sum_{k=1}^{K} \left(V_0^\star - V_0^{\pi^k} \right).$$

- Each term $V_0^\star - V_0^{\pi^k} \geq 0$ measures the sub-optimality gap of episode k.
- $R(K) = 0$ iff every episode plays an optimal policy.
- **Goal:** $R(K) = \widetilde{O}(\sqrt{K})$, so average sub-optimality $\to 0$.

Intuition

- Early episodes: explore (large $V_0^\star - V_0^{\pi^k}$, but learning fast)
- Later episodes: exploit (gaps shrink as estimates improve)
- $\sqrt{K}$ scaling: unavoidable cost of exploration

Online-to-Batch Conversion

Theorem (Online-to-Batch)

Suppose an algorithm guarantees $\mathbb{E}[R(K)] \leq B(K)$. Define $\widehat{\pi}$ by sampling uniformly at random from $\{\pi^1, \ldots, \pi^K\}$. Then

$$\mathbb{E}\left[V_0^\star - V_0^{\widehat{\pi}}\right] \leq \frac{B(K)}{K}.$$

If $B(K) = C\sqrt{K}$, then $K = O(C^2/\varepsilon^2)$ episodes achieve $\mathbb{E}[V_0^\star - V_0^{\widehat{\pi}}] \leq \varepsilon$.

Online-to-Batch Conversion

Theorem (Online-to-Batch)

Suppose an algorithm guarantees $\mathbb{E}[R(K)] \leq B(K)$. Define $\widehat{\pi}$ by sampling uniformly at random from $\{\pi^1, \ldots, \pi^K\}$. Then

$$\mathbb{E}\left[V_0^\star - V_0^{\widehat{\pi}}\right] \leq \frac{B(K)}{K}.$$

If $B(K) = C\sqrt{K}$, then $K = O(C^2/\varepsilon^2)$ episodes achieve $\mathbb{E}[V_0^\star - V_0^{\widehat{\pi}}] \leq \varepsilon$.

Proof. Let $U \sim \mathrm{Uniform}\{1, \ldots, K\}$ independent of the algorithm.

$$\mathbb{E}\left[V_0^\star - V_0^{\widehat{\pi}}\right] = \mathbb{E}\left[\frac{1}{K}\sum_{k=1}^{K}\left(V_0^\star - V_0^{\pi^k}\right)\right] = \frac{\mathbb{E}[R(K)]}{K} \leq \frac{B(K)}{K}.$$

Online-to-Batch Conversion

Theorem (Online-to-Batch)

Suppose an algorithm guarantees $\mathbb{E}[R(K)] \leq B(K)$. Define $\widehat{\pi}$ by sampling uniformly at random from $\{\pi^1, \ldots, \pi^K\}$. Then

$$\mathbb{E}\left[V_0^\star - V_0^{\widehat{\pi}}\right] \leq \frac{B(K)}{K}.$$

If $B(K) = C\sqrt{K}$, then $K = O(C^2/\varepsilon^2)$ episodes achieve $\mathbb{E}[V_0^\star - V_0^{\widehat{\pi}}] \leq \varepsilon$.

Proof. Let $U \sim \mathrm{Uniform}\{1, \ldots, K\}$ independent of the algorithm.

$$\mathbb{E}\left[V_0^\star - V_0^{\widehat{\pi}}\right] = \mathbb{E}\left[\frac{1}{K}\sum_{k=1}^{K}\left(V_0^\star - V_0^{\pi^k}\right)\right] = \frac{\mathbb{E}[R(K)]}{K} \leq \frac{B(K)}{K}.$$

By Markov's inequality, $\mathbb{P}\left(V_0^\star - V_0^{\widehat{\pi}} > 2B(K)/K\right) \leq 1/2$.
For $1-\delta$ confidence: run $\lceil \log_2(1/\delta) \rceil$ independent trials; return the empirically best policy. □

Online-to-Batch Conversion

Theorem (Online-to-Batch)

Suppose an algorithm guarantees $\mathbb{E}[R(K)] \leq B(K)$. Define $\widehat{\pi}$ by sampling uniformly at random from $\{\pi^1, \ldots, \pi^K\}$. Then

$$\mathbb{E}\left[V_0^\star - V_0^{\widehat{\pi}}\right] \leq \frac{B(K)}{K}.$$

If $B(K) = C\sqrt{K}$, then $K = O(C^2/\varepsilon^2)$ episodes achieve $\mathbb{E}[V_0^\star - V_0^{\widehat{\pi}}] \leq \varepsilon$.

Proof. Let $U \sim \mathrm{Uniform}\{1, \ldots, K\}$ independent of the algorithm.

$$\mathbb{E}\left[V_0^\star - V_0^{\widehat{\pi}}\right] = \mathbb{E}\left[\frac{1}{K}\sum_{k=1}^{K}\left(V_0^\star - V_0^{\pi^k}\right)\right] = \frac{\mathbb{E}[R(K)]}{K} \leq \frac{B(K)}{K}.$$

By Markov's inequality, $\mathbb{P}\left(V_0^\star - V_0^{\widehat{\pi}} > 2B(K)/K\right) \leq 1/2$.
For $1-\delta$ confidence: run $\lceil \log_2(1/\delta) \rceil$ independent trials; return the empirically best policy. □

Takeaway

Today's Plan

1. The Online RL Setting
2. PAC Framework & Hardness
3. Regret & Online-to-Batch
4. Multi-Armed Bandits
5. Contextual Bandits
6. UCB Value Iteration

Multi-Armed Bandit Setting

T rounds, N arms with unknown means $\mu_1, \ldots, \mu_N \in [0, 1]$.

At round t: pull arm i_t, observe reward $r_t \sim \mathrm{Dist}(\mu_{i_t})$, $r_t \in [0, 1]$.

Goal

Minimize regret against the best arm $i^\star = \arg\max_i \mu_i$:

$$R(T) \triangleq \sum_{t=1}^{T} \left(\mu_{i^\star} - \mu_{i_t}\right).$$

Multi-Armed Bandit Setting

T rounds, N arms with unknown means $\mu_1, \ldots, \mu_N \in [0, 1]$.

At round t: pull arm i_t, observe reward $r_t \sim \mathrm{Dist}(\mu_{i_t})$, $r_t \in [0, 1]$.

Goal

Minimize regret against the best arm $i^\star = \arg\max_i \mu_i$:

$$R(T) \triangleq \sum_{t=1}^{T} \left(\mu_{i^\star} - \mu_{i_t}\right).$$

Key quantities (parallel to MDP):

MAB	**MDP (preview)**
Round t	Episode k
Arm $i \in [N]$	State-action-stage (s, a, h)
Count $n_t(i)$	Count $n_h^k(s, a)$
True mean μ_i	True value $Q_h^\star(s, a)$

UCB Algorithm for MAB

UCB

For round $t = 1, 2, \ldots, T$**:**

1. Compute empirical means $\hat{\mu}_t(i) = \frac{1}{n_t(i)} \sum_{s \leq t:\, i_s = i} r_s$ and bonuses

$$b_t(i) \triangleq \sqrt{\frac{\log(2NT/\delta)}{2\, n_t(i)}}.$$

2. Pull arm $i_t = \arg\max_{i \in [N]} \; \hat{\mu}_t(i) + b_t(i)$.

UCB Algorithm for MAB

UCB

For round $t = 1, 2, \ldots, T$**:**

1. Compute empirical means $\hat{\mu}_t(i) = \frac{1}{n_t(i)} \sum_{s \le t:\, i_s = i} r_s$ and bonuses

$$b_t(i) \triangleq \sqrt{\frac{\log(2NT/\delta)}{2\, n_t(i)}}.$$

2. Pull arm $i_t = \arg\max_{i \in [N]} \; \hat{\mu}_t(i) + b_t(i)$.

Intuition: $\hat{\mu}_t(i) + b_t(i)$ is an optimistic upper bound on μ_i.

- Large $b_t(i)$: arm i is underexplored $\Rightarrow$ explore.
- Small $b_t(i)$: arm i is well-estimated $\Rightarrow$ exploit if $\hat{\mu}_t(i)$ is high.

UCB Algorithm for MAB

UCB

For round $t = 1, 2, \ldots, T$**:**

1. Compute empirical means $\hat{\mu}_t(i) = \frac{1}{n_t(i)} \sum_{s \le t:\, i_s = i} r_s$ and bonuses

$$b_t(i) \triangleq \sqrt{\frac{\log(2NT/\delta)}{2\, n_t(i)}}.$$

2. Pull arm $i_t = \arg\max_{i \in [N]} \; \hat{\mu}_t(i) + b_t(i)$.

Intuition: $\hat{\mu}_t(i) + b_t(i)$ is an optimistic upper bound on μ_i.

- Large $b_t(i)$: arm i is underexplored $\Rightarrow$ explore.
- Small $b_t(i)$: arm i is well-estimated $\Rightarrow$ exploit if $\hat{\mu}_t(i)$ is high.

Concentration event

$$\mathcal{E} \triangleq \bigcap_{t=1}^{T} \bigcap_{i=1}^{N} \left\{ |\hat{\mu}_t(i) - \mu_i| \le b_t(i) \right\}.$$

UCB Optimism & Regret Decomposition

Lemma (Optimism)

On $\mathcal{E}$, for all i and t:

$$\hat{\mu}_t(i) + b_t(i) \;\geq\; \mu_i.$$

In particular, the UCB of the chosen arm i_t upper-bounds $\mu_{i^\star}$:
$\hat{\mu}_t(i_t) + b_t(i_t) \geq \mu_{i^\star}$.

UCB Optimism & Regret Decomposition

Lemma (Optimism)

On $\mathcal{E}$, for all i and t:

$$\hat{\mu}_t(i) + b_t(i) \;\geq\; \mu_i.$$

In particular, the UCB of the chosen arm i_t upper-bounds $\mu_{i^\star}$:
$\hat{\mu}_t(i_t) + b_t(i_t) \geq \mu_{i^\star}$.

Proof: Direct from the definition of $\mathcal{E}$ (left inequality) and the fact that UCB picks the maximum index (so i_t's index $\geq$ $i^\star$'s index $\geq \mu_{i^\star}$). □

UCB Optimism & Regret Decomposition

Lemma (Optimism)

On $\mathcal{E}$, for all i and t:

$$\hat{\mu}_t(i) + b_t(i) \;\geq\; \mu_i.$$

In particular, the UCB of the chosen arm i_t upper-bounds $\mu_{i^\star}$:
$\hat{\mu}_t(i_t) + b_t(i_t) \geq \mu_{i^\star}$.

Proof: Direct from the definition of $\mathcal{E}$ (left inequality) and the fact that UCB picks the maximum index (so i_t's index $\geq$ $i^\star$'s index $\geq \mu_{i^\star}$). □

Regret decomposition: On $\mathcal{E}$, for each round t:

$$\begin{aligned}\mu_{i^\star} - \mu_{i_t} &\leq \left[\hat{\mu}_t(i_t) + b_t(i_t)\right] - \mu_{i_t} \\ &\leq \left[\hat{\mu}_t(i_t) + b_t(i_t)\right] - \left[\hat{\mu}_t(i_t) - b_t(i_t)\right] = 2\, b_t(i_t).\end{aligned}$$

UCB Optimism & Regret Decomposition

Lemma (Optimism)

On $\mathcal{E}$, for all i and t:

$$\hat{\mu}_t(i) + b_t(i) \;\geq\; \mu_i.$$

In particular, the UCB of the chosen arm i_t upper-bounds $\mu_{i^\star}$:
$\hat{\mu}_t(i_t) + b_t(i_t) \geq \mu_{i^\star}$.

Proof: Direct from the definition of $\mathcal{E}$ (left inequality) and the fact that UCB picks the maximum index (so i_t's index $\geq$ $i^\star$'s index $\geq \mu_{i^\star}$). □

Regret decomposition: On $\mathcal{E}$, for each round t:

$$\begin{aligned}\mu_{i^\star} - \mu_{i_t} &\leq \left[\hat{\mu}_t(i_t) + b_t(i_t)\right] - \mu_{i_t} \\ &\leq \left[\hat{\mu}_t(i_t) + b_t(i_t)\right] - \left[\hat{\mu}_t(i_t) - b_t(i_t)\right] = 2\, b_t(i_t).\end{aligned}$$

Summing over t:

$$R(T) \;\leq\; 2\sum_{t=1}^{T} b_t(i_t) = \sqrt{2\log(2NT/\delta)}\sum_{t=1}^{T}\frac{1}{\sqrt{n_t(i_t)}}.$$

UCB Regret Bound for MAB

Counting argument (same idea as MDP): for each arm i, the j-th pull contributes $1/\sqrt{j}$:

$$\sum_{t=1}^{T} \frac{1}{\sqrt{n_t(i_t)}} \leq \sum_{i=1}^{N} \sum_{j=1}^{n_T(i)} \frac{1}{\sqrt{j}} \leq 2 \sum_{i=1}^{N} \sqrt{n_T(i)} \leq 2\sqrt{N \cdot T}.$$

UCB Regret Bound for MAB

Counting argument (same idea as MDP): for each arm i, the j-th pull contributes $1/\sqrt{j}$:

$$\sum_{t=1}^{T} \frac{1}{\sqrt{n_t(i_t)}} \leq \sum_{i=1}^{N} \sum_{j=1}^{n_T(i)} \frac{1}{\sqrt{j}} \leq 2 \sum_{i=1}^{N} \sqrt{n_T(i)} \leq 2\sqrt{N \cdot T}.$$

Theorem (UCB Regret Bound)

With probability at least $1 - \delta$,

$$R(T) \leq 2\sqrt{2NT \log(2NT/\delta)}.$$

UCB Regret Bound for MAB

Counting argument (same idea as MDP): for each arm i, the j-th pull contributes $1/\sqrt{j}$:

$$\sum_{t=1}^{T} \frac{1}{\sqrt{n_t(i_t)}} \leq \sum_{i=1}^{N} \sum_{j=1}^{n_T(i)} \frac{1}{\sqrt{j}} \leq 2 \sum_{i=1}^{N} \sqrt{n_T(i)} \leq 2\sqrt{N \cdot T}.$$

Theorem (UCB Regret Bound)

With probability at least $1 - \delta$,

$$R(T) \leq 2\sqrt{2NT \log(2NT/\delta)}.$$

PAC consequence (online-to-batch): output arm $\hat{\imath}$ uniformly at random from $\{i_1, \ldots, i_T\}$.

$$\mathbb{E}[\mu_{i^\star} - \mu_{\hat{\imath}}] \leq \frac{\mathbb{E}[R(T)]}{T} \leq \frac{2\sqrt{2NT \log(\cdot)}}{T} = \frac{2\sqrt{2N \log(\cdot)}}{\sqrt{T}}.$$

Setting $T = O(N \log(N/\delta)/\varepsilon^2)$ gives an ε-optimal arm. *Same cost as uniform exploration!*

Today's Plan

1. The Online RL Setting
2. PAC Framework & Hardness
3. Regret & Online-to-Batch
4. Multi-Armed Bandits
5. Contextual Bandits
6. UCB Value Iteration

Contextual Bandit Setting

At round t: observe context $x_t \in \mathcal{X}$ (e.g. a user feature vector), choose $a_t \in \mathcal{A}$, observe reward.

Setting

- Finite context set $\mathcal{X}$, $X = |\mathcal{X}|$; action set $\mathcal{A}$, $A = |\mathcal{A}|$
- Unknown mean reward $\mu(x, a) = \mathbb{E}[r_t \mid x_t = x,\ a_t = a]$
- Optimal policy $\pi^\star(x) = \arg\max_a \mu(x, a)$
- Regret: $R(T) = \sum_{t=1}^{T} \big(\mu(x_t, \pi^\star(x_t)) - \mu(x_t, a_t)\big)$

Contextual Bandit Setting

At round t: observe context $x_t \in \mathcal{X}$ (e.g. a user feature vector), choose $a_t \in \mathcal{A}$, observe reward.

Setting

- Finite context set $\mathcal{X}$, $X = |\mathcal{X}|$; action set $\mathcal{A}$, $A = |\mathcal{A}|$
- Unknown mean reward $\mu(x, a) = \mathbb{E}[r_t \mid x_t = x,\ a_t = a]$
- Optimal policy $\pi^\star(x) = \arg\max_a \mu(x, a)$
- Regret: $R(T) = \sum_{t=1}^{T} \big(\mu(x_t, \pi^\star(x_t)) - \mu(x_t, a_t)\big)$

Key insight: treat each pair $(x, a) \in \mathcal{X} \times \mathcal{A}$ as a separate *arm*.
There are XA arms total; the context x_t tells us which subset of arms is relevant this round.

Contextual Bandit Setting

At round t: observe context $x_t \in \mathcal{X}$ (e.g. a user feature vector), choose $a_t \in \mathcal{A}$, observe reward.

Setting

- Finite context set $\mathcal{X}$, $X = |\mathcal{X}|$; action set $\mathcal{A}$, $A = |\mathcal{A}|$
- Unknown mean reward $\mu(x, a) = \mathbb{E}[r_t \mid x_t = x,\ a_t = a]$
- Optimal policy $\pi^\star(x) = \arg\max_a \mu(x, a)$
- Regret: $R(T) = \sum_{t=1}^{T} \big(\mu(x_t, \pi^\star(x_t)) - \mu(x_t, a_t)\big)$

Key insight: treat each pair $(x, a) \in \mathcal{X} \times \mathcal{A}$ as a separate *arm*.
There are XA arms total; the context x_t tells us which subset of arms is relevant this round.

Comparison to MAB

MAB: N arms, always choose from all N.
Contextual bandit: XA arms, but at round t only the A arms $\{(x_t, a) : a \in \mathcal{A}\}$ are

UCB for Contextual Bandits

Contextual UCB

For round $t = 1, 2, \ldots, T$**:**

1. Observe context x_t. Maintain per-pair counts $n_t(x, a)$, means $\hat{\mu}_t(x, a)$, and bonuses

$$b_t(x, a) \triangleq \sqrt{\frac{\log(2XAT/\delta)}{2\, n_t(x, a)}}.$$

2. Choose $a_t = \arg\max_{a \in \mathcal{A}} \ \hat{\mu}_t(x_t, a) + b_t(x_t, a)$.

UCB for Contextual Bandits

Contextual UCB

For round $t = 1, 2, \ldots, T$**:**

1. Observe context x_t. Maintain per-pair counts $n_t(x, a)$, means $\hat{\mu}_t(x, a)$, and bonuses
$$b_t(x, a) \triangleq \sqrt{\frac{\log(2XAT/\delta)}{2\, n_t(x, a)}}.$$
2. Choose $a_t = \arg\max_{a \in \mathcal{A}} \ \hat{\mu}_t(x_t, a) + b_t(x_t, a)$.

Identical structure to UCB: just index by (x, a) instead of i.

Concentration event

$$\mathcal{E} \triangleq \bigcap_{t,x,a} \left\{ |\hat{\mu}_t(x, a) - \mu(x, a)| \leq b_t(x, a) \right\}.$$

Hoeffding + union bound over XAT triples: $\mathbb{P}(\mathcal{E}) \geq 1 - \delta$.

Contextual UCB: Regret Bound

Regret decomposition (identical to MAB, index by (x, a)): on $\mathcal{E}$,

$$\mu(x_t, \pi^\star(x_t)) - \mu(x_t, a_t) \;\leq\; 2\, b_t(x_t, a_t).$$

Contextual UCB: Regret Bound

Regret decomposition (identical to MAB, index by (x, a)): on $\mathcal{E}$,

$$\mu(x_t, \pi^\star(x_t)) - \mu(x_t, a_t) \;\leq\; 2\, b_t(x_t, a_t).$$

Counting: $\sum_{x,a} n_T(x, a) = T$, so by the same argument as MAB:

$$\sum_{t=1}^{T} \frac{1}{\sqrt{n_t(x_t, a_t)}} \;\leq\; 2 \sum_{x,a} \sqrt{n_T(x, a)} \;\leq\; 2\sqrt{XA \cdot T}.$$

Contextual UCB: Regret Bound

Regret decomposition (identical to MAB, index by (x, a)): on $\mathcal{E}$,

$$\mu(x_t, \pi^\star(x_t)) - \mu(x_t, a_t) \;\leq\; 2\, b_t(x_t, a_t).$$

Counting: $\sum_{x,a} n_T(x, a) = T$, so by the same argument as MAB:

$$\sum_{t=1}^{T} \frac{1}{\sqrt{n_t(x_t, a_t)}} \;\leq\; 2 \sum_{x,a} \sqrt{n_T(x, a)} \;\leq\; 2\sqrt{XA \cdot T}.$$

Theorem (Contextual UCB Regret Bound)

With probability at least $1 - \delta$,

$$R(T) \;\leq\; 2\sqrt{2XAT \log(2XAT/\delta)}.$$

Contextual UCB: Regret Bound

Regret decomposition (identical to MAB, index by (x, a)): on $\mathcal{E}$,

$$\mu(x_t, \pi^\star(x_t)) - \mu(x_t, a_t) \;\le\; 2\, b_t(x_t, a_t).$$

Counting: $\sum_{x,a} n_T(x, a) = T$, so by the same argument as MAB:

$$\sum_{t=1}^{T} \frac{1}{\sqrt{n_t(x_t, a_t)}} \;\le\; 2 \sum_{x,a} \sqrt{n_T(x, a)} \;\le\; 2\sqrt{XA \cdot T}.$$

Theorem (Contextual UCB Regret Bound)

With probability at least $1 - \delta$,

$$R(T) \;\le\; 2\sqrt{2XAT \log(2XAT/\delta)}.$$

Observation: the result is exactly the MAB bound with $N \to XA$. The contexts add no extra cost beyond the increase in the number of effective arms.

From Contextual Bandits to MDPs

Natural idea: treat each (s, a, h) triple as an "arm".

	Contextual Bandit	**MDP**
Context	$x_t \in \mathcal{X}$	state $s_h \in \mathcal{S}$ at stage h
Action	$a_t \in \mathcal{A}$	$a_h \in \mathcal{A}$
# "arms"	XA	SAH
Target to estimate	$\mu(x, a)$	$Q_h^\star(s, a)$
Observable?	Yes — see r_t directly	No — $Q_h^\star$ involves future

From Contextual Bandits to MDPs

Natural idea: treat each (s, a, h) triple as an "arm".

	Contextual Bandit	**MDP**
Context	$x_t \in \mathcal{X}$	state $s_h \in \mathcal{S}$ at stage h
Action	$a_t \in \mathcal{A}$	$a_h \in \mathcal{A}$
# "arms"	XA	SAH
Target to estimate	$\mu(x, a)$	$Q_h^\star(s, a)$
Observable?	Yes — see r_t directly	No — $Q_h^\star$ involves future

The fundamental difficulty

In a contextual bandit we can estimate $\mu(x, a)$ directly from observed rewards.

In an MDP, $Q_h^\star(s, a) = r_h(s, a) + \mathbb{E}\left[\max_{a'} Q_{h+1}^\star(s', a')\right]$ depends on *future optimal play* — we cannot observe it directly.

Today's Plan

1. The Online RL Setting
2. PAC Framework & Hardness
3. Regret & Online-to-Batch
4. Multi-Armed Bandits
5. Contextual Bandits
6. UCB Value Iteration

Key Idea: Optimism

Principle of Optimism Under Uncertainty (OFU):

Optimism rule

At each episode, act as if the best plausible MDP (consistent with data so far) is the true one.

Why does this work?

- If optimistic estimate $\approx$ truth: we play a near-optimal policy. ✓
- If optimistic estimate $\gg$ truth: that (s, a, h) is underexplored $\Rightarrow$ we learn. ✓
- Either way, we make progress.

Implementation in UCB-VI:

- Build empirical MDP $\widehat{P}_h^k$ from past episodes $1, \ldots, k-1$.
- Add exploration bonus $b_h^k(s, a) \propto 1/\sqrt{n_h^k(s, a)}$ to each Bellman backup.
- Bonus is large for underexplored (s, a, h) and shrinks as data accumulates.

UCB-VI Algorithm

UCB-VI (Azar et al., 2017)

Input: horizon H, rewards $\{r_h\}$, confidence $\delta \in (0,1)$

For episode $k = 1, 2, \ldots, K$**:**

1. Form empirical kernels $\widehat{P}_h^k$ from episodes $1, \ldots, k-1$.
2. Set $\widehat{V}_{H+1}^k \equiv 0$ and bonuses $b_h^k(s,a) \triangleq H\sqrt{\dfrac{\log(2KHSA/\delta)}{2\, n_h^k(s,a)}}$.
3. For $h = H, H-1, \ldots, 1$, compute optimistic Q-values:

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

$$\widehat{V}_h^k(s) = \max_a\ \widehat{Q}_h^k(s,a).$$

4. Execute $\pi_h^k(s) = \arg\max_a\ \widehat{Q}_h^k(s,a)$ for one episode.

Concentration Events

The analysis relies on two high-probability events.

$\mathcal{E}_{\text{opt}}$ — Optimism event

$$\mathcal{E}_{\text{opt}} \triangleq \bigcap_{k,h,s,a} \left\{ \left| ((P_h - \widehat{P}_h^k) V_{h+1}^\star)(s,a) \right| \leq b_h^k(s,a) \right\}.$$

Concentration Events

The analysis relies on two high-probability events.

$\mathcal{E}_{\text{opt}}$ — Optimism event

$$\mathcal{E}_{\text{opt}} \triangleq \bigcap_{k,h,s,a} \left\{ \left| ((P_h - \widehat{P}_h^k) V_{h+1}^\star)(s,a) \right| \leq b_h^k(s,a) \right\}.$$

$\mathcal{E}_{\text{comp}}$ — Empirical-process event

$$\mathcal{E}_{\text{comp}} \triangleq \bigcap_{k,h,s,a} \left\{ \sup_{V \in [0,H]^S} \left| ((\widehat{P}_h^k - P_h) V)(s,a) \right| \leq H \sqrt{\frac{S \log(2KHSA/\delta)}{2\, n_h^k(s,a)}} \right\}.$$

Concentration Events

The analysis relies on two high-probability events.

$\mathcal{E}_{\text{opt}}$ — Optimism event

$$\mathcal{E}_{\text{opt}} \triangleq \bigcap_{k,h,s,a} \left\{ \left| ((P_h - \widehat{P}_h^k) V_{h+1}^\star)(s,a) \right| \leq b_h^k(s,a) \right\}.$$

$\mathcal{E}_{\text{comp}}$ — Empirical-process event

$$\mathcal{E}_{\text{comp}} \triangleq \bigcap_{k,h,s,a} \left\{ \sup_{V \in [0,H]^S} \left| ((\widehat{P}_h^k - P_h) V)(s,a) \right| \leq H \sqrt{\frac{S \log(2KHSA/\delta)}{2\, n_h^k(s,a)}} \right\}.$$

Lemma: $\mathbb{P}(\mathcal{E}_{\text{opt}}) \geq 1 - \delta$ and $\mathbb{P}(\mathcal{E}_{\text{comp}}) \geq 1 - \delta$. Hence $\mathbb{P}(\mathcal{E}_{\text{opt}} \cap \mathcal{E}_{\text{comp}}) \geq 1 - 2\delta$ by a union bound.

Why the Concentration Events Hold

Both proofs: Hoeffding on a fixed tuple, then union bound over all tuples.

$\mathcal{E}_{\text{opt}}$ — **fix** (k, h, s, a)**:** Define $X_i = \mathbf{1}\{(s_h^i, a_h^i) = (s, a)\}\big(\mathbb{E}_{P_h}[V_{h+1}^\star] - V_{h+1}^\star(s_{h+1}^i)\big)$.
Each X_i is a martingale difference with $|X_i| \leq H \cdot \mathbf{1}\{(s_h^i, a_h^i) = (s, a)\}$.

Why the Concentration Events Hold

Both proofs: Hoeffding on a fixed tuple, then union bound over all tuples.

$\mathcal{E}_{\text{opt}}$ — **fix** (k, h, s, a)**:** Define $X_i = \mathbf{1}\{(s_h^i, a_h^i) = (s, a)\}\big(\mathbb{E}_{P_h}[V_{h+1}^\star] - V_{h+1}^\star(s_{h+1}^i)\big)$.
Each X_i is a martingale difference with $|X_i| \le H \cdot \mathbf{1}\{(s_h^i, a_h^i) = (s, a)\}$.
Azuma–Hoeffding gives, after dividing by $n_h^k(s, a)$:

$$\left|((P_h - \widehat{P}_h^k)V_{h+1}^\star)(s, a)\right| \le H\sqrt{\frac{\log(2/\delta')}{2\, n_h^k(s, a)}}$$

with prob. $\ge 1 - \delta'$. Set $\delta' = \delta/(KHSA)$; union-bound over $KHSA$ tuples
$\Rightarrow \mathbb{P}(\mathcal{E}_{\text{opt}}) \ge 1 - \delta$.

Why the Concentration Events Hold

Both proofs: Hoeffding on a fixed tuple, then union bound over all tuples.

$\mathcal{E}_{\text{opt}}$ — **fix** (k, h, s, a)**:** Define $X_i = \mathbf{1}\{(s_h^i, a_h^i) = (s, a)\}\big(\mathbb{E}_{P_h}[V_{h+1}^\star] - V_{h+1}^\star(s_{h+1}^i)\big)$.
Each X_i is a martingale difference with $|X_i| \le H \cdot \mathbf{1}\{(s_h^i, a_h^i) = (s, a)\}$.
Azuma–Hoeffding gives, after dividing by $n_h^k(s, a)$:

$$\left|((P_h - \widehat{P}_h^k)V_{h+1}^\star)(s, a)\right| \le H\sqrt{\frac{\log(2/\delta')}{2\, n_h^k(s, a)}}$$

with prob. $\ge 1 - \delta'$. Set $\delta' = \delta/(KHSA)$; union-bound over $KHSA$ tuples
$\Rightarrow \mathbb{P}(\mathcal{E}_{\text{opt}}) \ge 1 - \delta$.

$\mathcal{E}_{\text{comp}}$ — **uniform bound over** $V \in [0, H]^S$**:** By linearity,
$\sup_{V \in [0,H]^S} |(\widehat{P}^k - P)V| = \max_{V \in \{0,H\}^S} |(\widehat{P}^k - P)V|$.

Why the Concentration Events Hold

Both proofs: Hoeffding on a fixed tuple, then union bound over all tuples.

$\mathcal{E}_{\text{opt}}$ — **fix** (k, h, s, a)**:** Define $X_i = \mathbf{1}\{(s_h^i, a_h^i) = (s, a)\}\big(\mathbb{E}_{P_h}[V_{h+1}^\star] - V_{h+1}^\star(s_{h+1}^i)\big)$.
Each X_i is a martingale difference with $|X_i| \le H \cdot \mathbf{1}\{(s_h^i, a_h^i) = (s, a)\}$.
Azuma–Hoeffding gives, after dividing by $n_h^k(s, a)$:

$$\left|((P_h - \widehat{P}_h^k)V_{h+1}^\star)(s, a)\right| \le H\sqrt{\frac{\log(2/\delta')}{2\, n_h^k(s, a)}}$$

with prob. $\ge 1 - \delta'$. Set $\delta' = \delta/(KHSA)$; union-bound over $KHSA$ tuples
$\Rightarrow \mathbb{P}(\mathcal{E}_{\text{opt}}) \ge 1-\delta$.

$\mathcal{E}_{\text{comp}}$ — **uniform bound over** $V \in [0, H]^S$**:** By linearity,
$\sup_{V\in[0,H]^S} |(\widehat{P}^k - P)V| = \max_{V\in\{0,H\}^S} |(\widehat{P}^k - P)V|$.
There are 2^S extreme points; apply Hoeffding to each and union-bound over them
$(\log 2^S = S \log 2)$:

$$\sup_{V\in[0,H]^S} \left|((\widehat{P}_h^k - P_h)V)(s, a)\right| \le H\sqrt{\frac{S\log(2/\delta')}{2\, n_h^k(s, a)}}.$$

Union-bound over $(k, h, s, a) \Rightarrow \mathbb{P}(\mathcal{E}_{\text{comp}}) \ge 1-\delta$.

Optimism Lemma

Lemma (Optimism)

On $\mathcal{E}_{\text{opt}}$, for all (k, h, s, a):

$$\widehat{Q}_h^k(s,a) \geq Q_h^\star(s,a) \qquad \text{and} \qquad \widehat{V}_h^k(s) \geq V_h^\star(s).$$

Optimism Lemma

Lemma (Optimism)

On $\mathcal{E}_{\text{opt}}$, for all (k, h, s, a):

$$\widehat{Q}_h^k(s,a) \geq Q_h^\star(s,a) \qquad \text{and} \qquad \widehat{V}_h^k(s) \geq V_h^\star(s).$$

Proof by backward induction on h.
Base case $h = H+1$: trivial since $\widehat{V}_{H+1}^k = V_{H+1}^\star = 0$.

Optimism Lemma

Lemma (Optimism)

On $\mathcal{E}_{\text{opt}}$, for all (k, h, s, a):

$$\widehat{Q}_h^k(s,a) \geq Q_h^\star(s,a) \qquad \text{and} \qquad \widehat{V}_h^k(s) \geq V_h^\star(s).$$

Proof by backward induction on h.
Base case $h = H+1$: trivial since $\widehat{V}_{H+1}^k = V_{H+1}^\star = 0$.

Inductive step: assume $\widehat{V}_{h+1}^k \geq V_{h+1}^\star$ pointwise. For any (s, a):

$$\begin{aligned} Q_h^\star(s,a) &= r_h(s,a) + (P_h V_{h+1}^\star)(s,a) \\ &= r_h(s,a) + (\widehat{P}_h^k V_{h+1}^\star)(s,a) + \underbrace{((P_h - \widehat{P}_h^k) V_{h+1}^\star)(s,a)}_{\leq b_h^k(s,a) \text{ on } \mathcal{E}_{\text{opt}}} \\ &\leq r_h(s,a) + (\widehat{P}_h^k \widehat{V}_{h+1}^k)(s,a) + b_h^k(s,a) \;\leq\; \widehat{Q}_h^k(s,a). \quad \square \end{aligned}$$

Optimism Lemma

Lemma (Optimism)

On $\mathcal{E}_{\text{opt}}$, for all (k, h, s, a):

$$\widehat{Q}_h^k(s,a) \geq Q_h^\star(s,a) \qquad \text{and} \qquad \widehat{V}_h^k(s) \geq V_h^\star(s).$$

Proof by backward induction on h.
Base case $h = H+1$: trivial since $\widehat{V}_{H+1}^k = V_{H+1}^\star = 0$.

Inductive step: assume $\widehat{V}_{h+1}^k \geq V_{h+1}^\star$ pointwise. For any (s, a):

$$\begin{aligned} Q_h^\star(s,a) &= r_h(s,a) + (P_h V_{h+1}^\star)(s,a) \\ &= r_h(s,a) + (\widehat{P}_h^k V_{h+1}^\star)(s,a) + \underbrace{((P_h - \widehat{P}_h^k) V_{h+1}^\star)(s,a)}_{\leq\, b_h^k(s,a) \text{ on } \mathcal{E}_{\text{opt}}} \\ &\leq r_h(s,a) + (\widehat{P}_h^k \widehat{V}_{h+1}^k)(s,a) + b_h^k(s,a) \;\leq\; \widehat{Q}_h^k(s,a). \quad \square \end{aligned}$$

Key consequence: on $\mathcal{E}_{\text{opt}}$,

$$V_0^\star - V_0^{\pi^k} \;\leq\; \mathbb{E}_{s_1 \sim \nu}\big[\widehat{V}_1^k(s_1) - V_1^{\pi^k}(s_1)\big] \;\triangleq\; \mathbb{E}\big[\Gamma_1^k(s_1^k)\big].$$

Basic Regret Bound for UCB-VI

Theorem (Basic Regret Bound, Azar et al. 2017)

With probability at least $1 - 2\delta$,

$$R(K) \leq H^2 S \sqrt{8AK \log(2KHSA/\delta)}.$$

- Scaling: $\widetilde{O}(H^2 S \sqrt{AK})$. Has an extra $\sqrt{S}$ — we will improve it.

Proof strategy

1. **Regret decomposition:** on $\mathcal{E}_{\text{opt}} \cap \mathcal{E}_{\text{comp}}$, bound $V_0^\star - V_0^{\pi^k}$ by a sum of bonuses.
2. **Counting argument:** $\sum_{k,h} \frac{1}{\sqrt{n_h^k(s_h^k, a_h^k)}} \leq 2H\sqrt{SAK}$.

Basic Bound: Regret Decomposition

Define $\Gamma_h^k \triangleq \widehat{V}_h^k - V_h^{\pi^k} \geq 0$ (optimism gap).
On $\mathcal{E}_{\text{opt}}$: $V_0^\star - V_0^{\pi^k} \leq \mathbb{E}[\Gamma_1^k(s_1^k)]$.

Basic Bound: Regret Decomposition

Define $\Gamma_h^k \triangleq \widehat{V}_h^k - V_h^{\pi^k} \geq 0$ (optimism gap).
On $\mathcal{E}_{\text{opt}}$: $V_0^\star - V_0^{\pi^k} \leq \mathbb{E}[\Gamma_1^k(s_1^k)]$.

One step of the unrolling (Bellman equations for $\widehat{V}^k$ and V^{π^k}):

$$\mathbb{E}[\Gamma_1^k(s_1^k)] = \mathbb{E}\left[b_1^k(s_1^k, a_1^k) + ((\widehat{P}_1^k - P_1)\widehat{V}_2^k)(s_1^k, a_1^k)\right] + \mathbb{E}[\Gamma_2^k(s_2^k)].$$

Basic Bound: Regret Decomposition

Define $\Gamma_h^k \triangleq \widehat{V}_h^k - V_h^{\pi^k} \geq 0$ (optimism gap).
On $\mathcal{E}_{\text{opt}}$: $V_0^\star - V_0^{\pi^k} \leq \mathbb{E}[\Gamma_1^k(s_1^k)]$.

One step of the unrolling (Bellman equations for $\widehat{V}^k$ and V^{π^k}):

$$\mathbb{E}[\Gamma_1^k(s_1^k)] = \mathbb{E}\left[b_1^k(s_1^k, a_1^k) + ((\widehat{P}_1^k - P_1)\widehat{V}_2^k)(s_1^k, a_1^k)\right] + \mathbb{E}[\Gamma_2^k(s_2^k)].$$

Iterating from $h = 1$ to H (using $\Gamma_{H+1}^k \equiv 0$):

$$V_0^\star - V_0^{\pi^k} \leq \sum_{h=1}^{H} \mathbb{E}\left[b_h^k(s_h^k, a_h^k) + ((\widehat{P}_h^k - P_h)\widehat{V}_{h+1}^k)(s_h^k, a_h^k)\right].$$

Basic Bound: Regret Decomposition

Define $\Gamma_h^k \triangleq \widehat{V}_h^k - V_h^{\pi^k} \geq 0$ (optimism gap).
On $\mathcal{E}_{\text{opt}}$: $V_0^\star - V_0^{\pi^k} \leq \mathbb{E}[\Gamma_1^k(s_1^k)]$.

One step of the unrolling (Bellman equations for $\widehat{V}^k$ and V^{π^k}):

$$\mathbb{E}[\Gamma_1^k(s_1^k)] = \mathbb{E}\left[b_1^k(s_1^k, a_1^k) + ((\widehat{P}_1^k - P_1)\widehat{V}_2^k)(s_1^k, a_1^k)\right] + \mathbb{E}[\Gamma_2^k(s_2^k)].$$

Iterating from $h = 1$ to H (using $\Gamma_{H+1}^k \equiv 0$):

$$V_0^\star - V_0^{\pi^k} \leq \sum_{h=1}^{H} \mathbb{E}\left[b_h^k(s_h^k, a_h^k) + ((\widehat{P}_h^k - P_h)\widehat{V}_{h+1}^k)(s_h^k, a_h^k)\right].$$

On $\mathcal{E}_{\text{comp}}$ the empirical-process term is bounded by $b_h^k \cdot \sqrt{S}$; both terms combine to:

$$V_0^\star - V_0^{\pi^k} \leq \sum_{h=1}^{H} \mathbb{E}\left[H\sqrt{\frac{2S\log(2KHSA/\delta)}{n_h^k(s_h^k, a_h^k)}}\right].$$

Basic Bound: Counting Argument

It remains to bound $\sum_{k=1}^{K} \sum_{h=1}^{H} \frac{1}{\sqrt{n_h^k(s_h^k, a_h^k)}}$.

Basic Bound: Counting Argument

It remains to bound $\sum_{k=1}^{K} \sum_{h=1}^{H} \frac{1}{\sqrt{n_h^k(s_h^k, a_h^k)}}$.

Fix h and (s, a). The i-th visit contributes $1/\sqrt{i}$ where i is the running count:

$$\sum_{k=1}^{K} \frac{\mathbf{1}\{(s_h^k, a_h^k) = (s, a)\}}{\sqrt{n_h^k(s, a)}} \leq \sum_{i=1}^{n_h^{K+1}(s,a)} \frac{1}{\sqrt{i}} \leq 2\sqrt{n_h^{K+1}(s, a)}.$$

Basic Bound: Counting Argument

It remains to bound $\sum_{k=1}^{K}\sum_{h=1}^{H}\frac{1}{\sqrt{n_h^k(s_h^k,a_h^k)}}$.

Fix h and (s,a). The i-th visit contributes $1/\sqrt{i}$ where i is the running count:

$$\sum_{k=1}^{K}\frac{\mathbf{1}\{(s_h^k,a_h^k)=(s,a)\}}{\sqrt{n_h^k(s,a)}} \;\leq\; \sum_{i=1}^{n_h^{K+1}(s,a)}\frac{1}{\sqrt{i}} \;\leq\; 2\sqrt{n_h^{K+1}(s,a)}.$$

Sum over (s,a) and apply Cauchy-Schwarz ($\sum_i \sqrt{x_i} \leq \sqrt{N\sum_i x_i}$):

$$\sum_{k=1}^{K}\frac{1}{\sqrt{n_h^k(s_h^k,a_h^k)}} \;\leq\; 2\sum_{s,a}\sqrt{n_h^{K+1}(s,a)} \;\leq\; 2\sqrt{SA\cdot K} \qquad (\text{since } \textstyle\sum_{s,a} n_h^{K+1}(s,a) = K$$

Basic Bound: Counting Argument

It remains to bound $\sum_{k=1}^{K}\sum_{h=1}^{H}\frac{1}{\sqrt{n_h^k(s_h^k,a_h^k)}}$.

Fix h and (s,a). The i-th visit contributes $1/\sqrt{i}$ where i is the running count:

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

Sum over (s,a) and apply Cauchy-Schwarz ($\sum_i \sqrt{x_i} \le \sqrt{N\sum_i x_i}$):

$$\sum_{k=1}^{K}\frac{1}{\sqrt{n_h^k(s_h^k,a_h^k)}} \;\le\; 2\sum_{s,a}\sqrt{n_h^{K+1}(s,a)} \;\le\; 2\sqrt{SA\cdot K} \qquad (\text{since } \textstyle\sum_{s,a} n_h^{K+1}(s,a) = K$$

Sum over $h\in[H]$: $\sum_{k,h}\frac{1}{\sqrt{n_h^k}} \le 2H\sqrt{SAK}$.

Plugging in: $R(K) \le H\sqrt{2S\log(\cdot)}\cdot 2H\sqrt{SAK} = H^2S\sqrt{8AK\log(\cdot)}$. □

Can We Do Better? A Sharper Analysis

Where did we lose? The event $\mathcal{E}_{\text{comp}}$ bounds $\sup_{V \in [0,H]^S} |(\widehat{P}^k - P)V|$ — a uniform bound that costs $\sqrt{S}$.

Key insight: in the regret decomposition, the empirical-process term always involves $f = \widehat{V}^k_{h+1} - V^\star_{h+1}$, a *specific* function we can analyze more carefully.

Improved bound

$$R(K) = \widetilde{O}\left(H^2\sqrt{SAK} + H^3S^2A\right).$$

Leading term $H^2\sqrt{SAK}$ matches the minimax lower bound up to log factors.

Tool: Bernstein inequality applied coordinatewise to $(P_h - \widehat{P}^k_h)f$.

The Key Bernstein Estimate

Key Estimate

With probability $\geq 1-\delta$, for all (h, s, a, k) and all $f : \mathcal{S} \to [0, H]$, writing $n = n_h^k(s, a)$:

$$\left|((P_h - \widehat{P}_h^k)f)(s, a)\right| \leq \frac{2H^2 S \log(2S/\delta)}{n} + \frac{1}{H}(P_h f)(s, a).$$

The Key Bernstein Estimate

Key Estimate

With probability $\geq 1-\delta$, for all (h, s, a, k) and all $f : \mathcal{S} \to [0, H]$, writing $n = n_h^k(s, a)$:

$$\left|((P_h - \widehat{P}_h^k)f)(s,a)\right| \leq \frac{2H^2 S \log(2S/\delta)}{n} + \frac{1}{H}(P_h f)(s,a).$$

Step 1 — Coordinatewise Bernstein (Bernstein's inequality on each coordinate s'):

$$\left|((P_h - \widehat{P}_h^k)f)(s,a)\right| \leq \sum_{s'} f(s') \left(\sqrt{\frac{2P_h(s'|s,a)\log(2S/\delta)}{n}} + \frac{2\log(2S/\delta)}{3n} \right).$$

The Key Bernstein Estimate

Key Estimate

With probability $\geq 1-\delta$, for all (h, s, a, k) and all $f : \mathcal{S} \to [0, H]$, writing $n = n_h^k(s, a)$:

$$\left|((P_h - \widehat{P}_h^k)f)(s,a)\right| \leq \frac{2H^2 S \log(2S/\delta)}{n} + \frac{1}{H}(P_h f)(s,a).$$

Step 1 — Coordinatewise Bernstein (Bernstein's inequality on each coordinate s'):

$$\left|((P_h - \widehat{P}_h^k)f)(s,a)\right| \leq \sum_{s'} f(s') \left(\sqrt{\frac{2P_h(s'|s,a)\log(2S/\delta)}{n}} + \frac{2\log(2S/\delta)}{3n} \right).$$

Step 2 — Cauchy-Schwarz on the $\sqrt{\cdot}$ sum ($\sum_i x_i y_i \leq \sqrt{\sum x_i^2}\sqrt{\sum y_i^2}$):

$$\leq \frac{2HS\log(\cdot)}{n} + \sqrt{\frac{2HS\log(\cdot)}{n} \cdot (P_h f)(s,a)}.$$

The Key Bernstein Estimate

Key Estimate

With probability $\geq 1-\delta$, for all (h, s, a, k) and all $f : \mathcal{S} \to [0, H]$, writing $n = n_h^k(s, a)$:

$$\left|((P_h - \widehat{P}_h^k)f)(s,a)\right| \;\leq\; \frac{2H^2 S \log(2S/\delta)}{n} + \frac{1}{H}(P_h f)(s,a).$$

Step 1 — Coordinatewise Bernstein (Bernstein's inequality on each coordinate s'):

$$\left|((P_h - \widehat{P}_h^k)f)(s,a)\right| \leq \sum_{s'} f(s') \left(\sqrt{\frac{2P_h(s'|s,a)\log(2S/\delta)}{n}} + \frac{2\log(2S/\delta)}{3n} \right).$$

Step 2 — Cauchy-Schwarz on the $\sqrt{\cdot}$ sum ($\sum_i x_i y_i \leq \sqrt{\sum x_i^2}\sqrt{\sum y_i^2}$):

$$\leq \frac{2HS\log(\cdot)}{n} + \sqrt{\frac{2HS\log(\cdot)}{n} \cdot (P_h f)(s,a)}.$$

Step 3 — AM-GM ($2\sqrt{UV} \leq V/H + HU$ with $U = \frac{2HS\log}{n}$, $V = P_h f$):

Improved Bound: One-Step Recursion

Apply the Bernstein estimate to $f = \widehat{V}^k_{h+1} - V^\star_{h+1} \in [0, H]$ (valid on $\mathcal{E}_{\text{opt}}$).

Improved Bound: One-Step Recursion

Apply the Bernstein estimate to $f = \widehat{V}^k_{h+1} - V^\star_{h+1} \in [0, H]$ (valid on $\mathcal{E}_{\text{opt}}$).

One step of the Γ^k_h recursion becomes:

$$\Gamma^k_h(s) \;\le\; \underbrace{H\sqrt{\frac{2\log(\cdot)}{n^k_h}}}_{\text{original bonus}} + \underbrace{\frac{2H^2 S \log(\cdot)}{n^k_h}}_{\text{Bernstein correction}} + \underbrace{\left(1 + \tfrac{1}{H}\right)(P_h \Gamma^k_{h+1})(s, \pi^k_h(s))}_{\text{inflated propagation}}.$$

Improved Bound: One-Step Recursion

Apply the Bernstein estimate to $f = \widehat{V}^k_{h+1} - V^\star_{h+1} \in [0, H]$ (valid on $\mathcal{E}_{\text{opt}}$).

One step of the Γ^k_h recursion becomes:

$$\Gamma^k_h(s) \;\le\; \underbrace{H\sqrt{\frac{2\log(\cdot)}{n^k_h}}}_{\text{original bonus}} + \underbrace{\frac{2H^2 S\log(\cdot)}{n^k_h}}_{\text{Bernstein correction}} + \underbrace{\left(1+\tfrac{1}{H}\right)(P_h\Gamma^k_{h+1})(s,\pi^k_h(s))}_{\text{inflated propagation}}.$$

The $1/H$ term from the Bernstein estimate inflates the propagation factor from 1 to $(1 + 1/H)$. This is the key trade-off: a tighter concentration bound costs a slightly larger recursion factor.

Improved Bound: Unrolling the Recursion

Let $\beta_h^k \triangleq H\sqrt{2\log(\cdot)/n_h^k} + 2H^2 S \log(\cdot)/n_h^k$ (per-step error).

Unrolling $\Gamma_h^k \leq \beta_h^k + (1 + \frac{1}{H})(P_h \Gamma_{h+1}^k)$ from $h = 1$ to H:

Improved Bound: Unrolling the Recursion

Let $\beta_h^k \triangleq H\sqrt{2\log(\cdot)/n_h^k} + 2H^2 S \log(\cdot)/n_h^k$ (per-step error).

Unrolling $\Gamma_h^k \leq \beta_h^k + (1 + \frac{1}{H})(P_h \Gamma_{h+1}^k)$ from $h = 1$ to H:

$$\begin{aligned}
\mathbb{E}[\Gamma_1^k] &\leq \mathbb{E}[\beta_1^k] + \left(1 + \tfrac{1}{H}\right)\mathbb{E}[\Gamma_2^k] \\
&\leq \mathbb{E}[\beta_1^k] + \left(1 + \tfrac{1}{H}\right)\mathbb{E}[\beta_2^k] + \left(1 + \tfrac{1}{H}\right)^2 \mathbb{E}[\Gamma_3^k] \\
&\vdots \\
&\leq \sum_{h=1}^{H} \left(1 + \tfrac{1}{H}\right)^{h-1} \mathbb{E}[\beta_h^k].
\end{aligned}$$

Improved Bound: Unrolling the Recursion

Let $\beta_h^k \triangleq H\sqrt{2\log(\cdot)/n_h^k} + 2H^2 S \log(\cdot)/n_h^k$ (per-step error).

Unrolling $\Gamma_h^k \leq \beta_h^k + (1 + \frac{1}{H})(P_h \Gamma_{h+1}^k)$ from $h = 1$ to H:

$$\begin{aligned}
\mathbb{E}[\Gamma_1^k] &\leq \mathbb{E}[\beta_1^k] + \left(1 + \tfrac{1}{H}\right)\mathbb{E}[\Gamma_2^k] \\
&\leq \mathbb{E}[\beta_1^k] + \left(1 + \tfrac{1}{H}\right)\mathbb{E}[\beta_2^k] + \left(1 + \tfrac{1}{H}\right)^2 \mathbb{E}[\Gamma_3^k] \\
&\vdots \\
&\leq \sum_{h=1}^{H} \left(1 + \tfrac{1}{H}\right)^{h-1} \mathbb{E}[\beta_h^k].
\end{aligned}$$

Since $(1 + 1/H)^{h-1} \leq (1 + 1/H)^H \leq e$ for all $h \leq H$:

$$\mathbb{E}[\Gamma_1^k] \;\leq\; e \sum_{h=1}^{H} \mathbb{E}[\beta_h^k].$$

Improved Regret Bound

Two sums to control after summing over k and h:

Sum 1 ($1/\sqrt{n}$ terms, as before):

$$\sum_{k=1}^{K} \sum_{h=1}^{H} \frac{1}{\sqrt{n_h^k(s_h^k, a_h^k)}} \leq 2H\sqrt{SAK}.$$

Improved Regret Bound

Two sums to control after summing over k and h:

Sum 1 ($1/\sqrt{n}$ terms, as before):

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\frac{1}{\sqrt{n_h^k(s_h^k,a_h^k)}} \leq 2H\sqrt{SAK}.$$

Sum 2 ($1/n$ terms, harmonic series):

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\frac{1}{n_h^k(s_h^k,a_h^k)} = \sum_{h,s,a}\sum_{i=1}^{n_h^{K+1}(s,a)}\frac{1}{i} \leq 2SAH\log K.$$

Improved Regret Bound

Two sums to control after summing over k and h:

Sum 1 ($1/\sqrt{n}$ terms, as before):

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\frac{1}{\sqrt{n_h^k(s_h^k,a_h^k)}} \;\leq\; 2H\sqrt{SAK}.$$

Sum 2 ($1/n$ terms, harmonic series):

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\frac{1}{n_h^k(s_h^k,a_h^k)} = \sum_{h,s,a}\sum_{i=1}^{n_h^{K+1}(s,a)}\frac{1}{i} \;\leq\; 2SAH\log K.$$

Theorem (Improved Regret Bound)

With probability at least $1-\delta$,

$$R(K) \;\leq\; H^2\sqrt{8e^2SAK\log(2KHSA/\delta)} \;+\; 4e\,H^3S^2A\log(K)\log(2KHS^2A/\delta).$$

In particular, $R(K) = \widetilde{O}(H^2\sqrt{SAK} + H^3S^2A)$.

Summary

Objective	UCB-VI rate	PAC via O-t-B
Regret (basic)	$\widetilde{O}(H^2 S\sqrt{AK})$	$\widetilde{O}(H^4 S^2 A/\varepsilon^2)$
Regret (improved)	$\widetilde{O}(H^2\sqrt{SAK})$	$\widetilde{O}(H^4 SA/\varepsilon^2)$

Big takeaways

- **Optimism (OFU)** drives exploration automatically — visit what you don't know.
- **Online-to-batch**: any $\widetilde{O}(\sqrt{K})$ regret bound gives a PAC algorithm for free.
- **Bernstein** $>$ **Hoeffding**: variance-aware bounds save a $\sqrt{S}$ factor in regret.