CSE 542: Statistical Reinforcement Learning

Lecture 4: Online Reinforcement Learning

Kevin Jamieson

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

Outline

Today's Plan

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

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

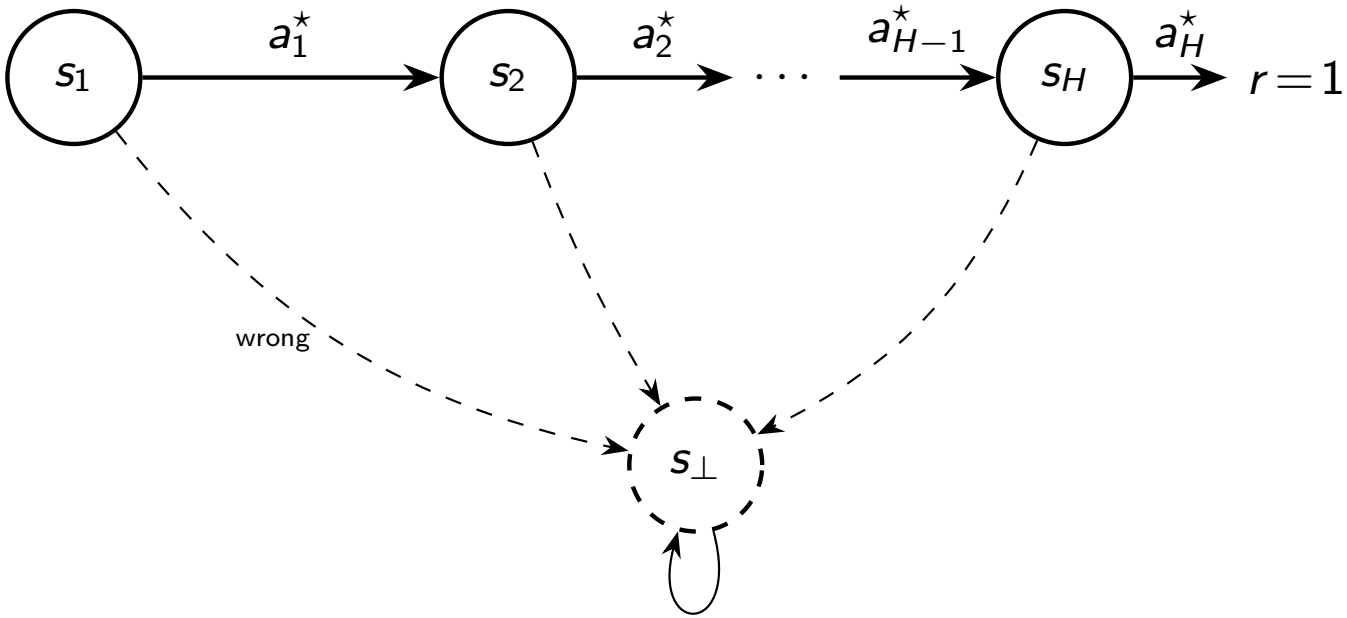
- States: $\mathcal{S} = \{s_1, \ldots, s_H, s_\perp\}$ ($s_\perp$ = absorbing dead state)
- Actions: $|\mathcal{A}| = A$; at each stage h there is one correct action $a_h^\star$
- Transitions from live state s_h:

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

- 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

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. $\square$

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

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

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

UCB-VI: The Algorithm

Key idea: run value iteration on the empirical MDP, but add an optimistic bonus to each Bellman backup to encourage exploration.

UCB-VI (Azar et al., 2017)

Input: horizon H, rewards $\{r_h\}_h$, confidence parameter δ.

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

1. Form empirical kernels $\widehat{P}_h^k$ from episodes $1, \ldots, k-1$. Set $\widehat{V}_{H+1}^k \equiv 0$.
2. For $h = H, H-1, \ldots, 1$ and all (s, a), set

$$b_h^k(s,a) := H\sqrt{\frac{\log(2KHSA/\delta)}{2\, n_h^k(s,a)}},$$

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

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

3. Let $\pi_h^k(s) \in \arg\max_a \widehat{Q}_h^k(s,a)$ and execute π^k.

Optimism vs. Pessimism

	Offline PEVI	**Online UCB-VI**
Bonus direction	Subtract $b_h(s, a)$	Add $b_h^k(s, a)$
Principle	**Pessimistic**: avoid uncertain	**Optimistic**: seek uncertain
Effect	Stay in-distribution	Explore undervisited (s, a, h)

Optimism vs. Pessimism

	Offline PEVI	**Online UCB-VI**
Bonus direction	Subtract $b_h(s,a)$	Add $b_h^k(s,a)$
Principle	**Pessimistic**: avoid uncertain	**Optimistic**: seek uncertain
Effect	Stay in-distribution	Explore undervisited (s,a,h)

The optimism principle

UCB-VI picks actions that maximize $\widehat{Q}_h^k(s,a)$. A large bonus b_h^k inflates the index of undervisited pairs. So the algorithm explores (s,a) either because:

- it *appears* good (exploit the estimate), or
- it remains *underexplored* (the bonus is still large).

In either case, cumulative error is controlled.

Concentration: The Optimism Event

Define the "good event" for certifying optimism:

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

Lemma (Transition deviation)

For any *fixed* $V : \mathcal{S} \to [0, H]$ and any (h, s, a, k):

$$\mathbb{P}\left(\left| ((P_h - \widehat{P}_h^k) V)(s,a) \right| \le H \sqrt{\frac{\log(2/\delta)}{2\, n_h^k(s,a)}} \right) \ge 1 - \delta.$$

Consequently, $\mathbb{P}(\mathcal{E}_{\text{opt}}) \ge 1 - \delta$.

Concentration: The Optimism Event

Define the "good event" for certifying optimism:

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

Lemma (Transition deviation)

For any *fixed* $V : \mathcal{S} \to [0, H]$ and any (h, s, a, k):

$$\mathbb{P}\left(\left|((P_h - \widehat{P}_h^k) V)(s,a)\right| \le H \sqrt{\frac{\log(2/\delta)}{2\, n_h^k(s,a)}} \right) \ge 1 - \delta.$$

Consequently, $\mathbb{P}(\mathcal{E}_{\text{opt}}) \ge 1 - \delta$.

Proof sketch: The i-th sample contributes to $((\widehat{P}_h^k - P_h)V)(s,a)$ only when $(s_h^i, a_h^i) = (s,a)$. Conditional on visiting (s,a), the increment is a bounded mean-zero martingale difference. Azuma–Hoeffding gives a single-pair bound; a union bound over all (h, s, a, k) replaces δ with $\delta/(KHSA)$.

Lemma: Optimism

Lemma (Optimism)

On $\mathcal{E}_{\text{opt}}$, for all episodes k, stages h, states s, actions a:

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

Lemma: Optimism

Lemma (Optimism)

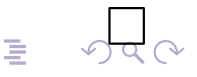
On $\mathcal{E}_{\text{opt}}$, for all episodes k, stages h, states s, actions a:

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

Proof (backward induction on h):
Base ($h = H+1$): both sides are zero. ✓
Inductive step: Assume $\widehat{V}_{h+1}^k \geq V_{h+1}^\star$. If $\widehat{Q}_h^k(s,a) = H$, then trivially $\geq Q_h^\star(s,a)$. Otherwise:

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

Taking a max over a gives $\widehat{V}_h^k(s) \geq V_h^\star(s)$.



Regret-to-Surplus Certificate

Lemma (Certificate)

Define the *surplus* at episode k and stage h:

$$\varepsilon_h^k(s,a) := \widehat{Q}_h^k(s,a) - r_h(s,a) - (P_h \widehat{V}_{h+1}^k)(s,a).$$

For every k, h, s, with $a = \pi_h^k(s)$ we have for $\Gamma_h^k := \widehat{V}_h^k - V_h^{\pi^k}$

$$\Gamma_h^k(s) = \varepsilon_h^k(s,a) + (P_h \Gamma_{h+1}^k)(s,a).$$

On $\mathcal{E}_{\mathrm{opt}}$, for every episode k: $V_0^\star - V_0^{\pi^k} \le \sum_{h=1}^H \mathbb{E}\left[\varepsilon_h^k(s_h^k, a_h^k)\right]$.

Regret-to-Surplus Certificate

Lemma (Certificate)

Define the *surplus* at episode k and stage h:

$$\varepsilon_h^k(s,a) := \widehat{Q}_h^k(s,a) - r_h(s,a) - (P_h \widehat{V}_{h+1}^k)(s,a).$$

For every k, h, s, with $a = \pi_h^k(s)$ we have for $\Gamma_h^k := \widehat{V}_h^k - V_h^{\pi^k}$

$$\Gamma_h^k(s) = \varepsilon_h^k(s,a) + (P_h \Gamma_{h+1}^k)(s,a).$$

On $\mathcal{E}_{\text{opt}}$, for every episode k: $V_0^\star - V_0^{\pi^k} \le \sum_{h=1}^H \mathbb{E}\left[\varepsilon_h^k(s_h^k, a_h^k)\right]$.

Proof sketch: Set $\Gamma_h^k := \widehat{V}_h^k - V_h^{\pi^k}$. At the visited state s_h^k with $a_h^k = \pi_h^k(s_h^k)$:

$$\begin{aligned}
\Gamma_h^k(s_h^k) = \widehat{V}_h^k(s_h^k) - V_h^{\pi^k}(s_h^k) &= \widehat{Q}_h^k(s_h^k, a_h^k) - r_h(s_h^k, a_h^k) - P_h V_{h+1}^{\pi^k}(s_h^k, a_h^k) \\
&= \varepsilon_h^k(s_h^k, a_h^k) + P_h(\widehat{V}_{h+1}^k - V_{h+1}^{\pi^k})(s_h^k, a_h^k) \\
&= \varepsilon_h^k(s_h^k, a_h^k) + (P_h \Gamma_{h+1}^k)(s_h^k, a_h^k) \\
&= \varepsilon_h^k(s_h^k, a_h^k) + \mathbb{E}[\Gamma_{h+1}^k(s_{h+1}^k) | s_h^k, a_h^k].
\end{aligned}$$

By optimism, $\widehat{V}_0^k \ge V_0^\star$, so $V_0^\star - V_0^{\pi^k} \le \widehat{V}_0^k - V_0^{\pi^k} = \mathbb{E}[\Gamma_1^k(s_1^k)] \le \sum_h \mathbb{E}[\varepsilon_h^k]$.

Bounding the Surplus: The Complexity Event

From the definitions: $\varepsilon_h^k(s,a) = b_h^k(s,a) + ((\widehat{P}_h^k - P_h)\widehat{V}_{h+1}^k)(s,a)$.

Problem: $\widehat{V}_{h+1}^k$ and $\widehat{P}_h^k$ share training data, so the transition deviation lemma (which requires a *fixed* function) does not apply directly.

Bounding the Surplus: The Complexity Event

From the definitions: $\varepsilon_h^k(s,a) = b_h^k(s,a) + ((\widehat{P}_h^k - P_h)\widehat{V}_{h+1}^k)(s,a)$.

Problem: $\widehat{V}_{h+1}^k$ and $\widehat{P}_h^k$ share training data, so the transition deviation lemma (which requires a *fixed* function) does not apply directly.

Complexity event $\mathcal{E}_{\text{comp}}$

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

$\mathbb{P}(\mathcal{E}_{\text{comp}}) \geq 1 - \delta.$

Bounding the Surplus: The Complexity Event

From the definitions: $\varepsilon_h^k(s,a) = b_h^k(s,a) + ((\widehat{P}_h^k - P_h)\widehat{V}_{h+1}^k)(s,a)$.

Problem: $\widehat{V}_{h+1}^k$ and $\widehat{P}_h^k$ share training data, so the transition deviation lemma (which requires a *fixed* function) does not apply directly.

Complexity event $\mathcal{E}_{\text{comp}}$

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

$\mathbb{P}(\mathcal{E}_{\text{comp}}) \geq 1 - \delta.$

Why? $V \mapsto ((\widehat{P}_h^k - P_h)V)(s,a)$ is linear, so its supremum over the hypercube $[0,H]^S$ is attained at a vertex $V \in \{0,H\}^S$. Apply the deviation lemma to each of the 2^S vertices, union bound, then union bound over (h,s,a,k) — gaining $\sqrt{S}$ from $\log(2^S) = S\log 2$.

Bounding the Surplus: The Complexity Event

From the definitions: $\varepsilon_h^k(s,a) = b_h^k(s,a) + ((\widehat{P}_h^k - P_h)\widehat{V}_{h+1}^k)(s,a)$.

Problem: $\widehat{V}_{h+1}^k$ and $\widehat{P}_h^k$ share training data, so the transition deviation lemma (which requires a *fixed* function) does not apply directly.

Complexity event $\mathcal{E}_{\text{comp}}$

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

$\mathbb{P}(\mathcal{E}_{\text{comp}}) \ge 1 - \delta.$

Why? $V \mapsto ((\widehat{P}_h^k - P_h)V)(s,a)$ is linear, so its supremum over the hypercube $[0,H]^S$ is attained at a vertex $V \in \{0,H\}^S$. Apply the deviation lemma to each of the 2^S vertices, union bound, then union bound over (h,s,a,k) — gaining $\sqrt{S}$ from $\log(2^S) = S\log 2$.

On $\mathcal{E}_{\text{opt}} \cap \mathcal{E}_{\text{comp}}$, the surplus satisfies $\varepsilon_h^k(s,a) \lesssim H\sqrt{S\log(\cdot)/n_h^k(s,a)}$.

Basic Regret Bound

Theorem (Basic regret bound for UCB-VI)

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

$$\sum_{k=1}^{K} \left(V_0^\star - V_0^{\pi^k} \right) \leq H^2 S \sqrt{8AK \log(2KHSA/\delta)}.$$

Basic Regret Bound

Theorem (Basic regret bound for UCB-VI)

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

$$\sum_{k=1}^{K} (V_0^\star - V_0^{\pi^k}) \leq H^2 S \sqrt{8AK \log(2KHSA/\delta)}.$$

Proof sketch: On $\mathcal{E}_{\text{opt}} \cap \mathcal{E}_{\text{comp}}$, combine the certificate with the surplus bound:

$$\sum_{k=1}^{K} (V_0^\star - V_0^{\pi^k}) \leq \sum_{k=1}^{K} \sum_{h=1}^{H} \mathbb{E}[\varepsilon_h^k(s_h^k, a_h^k)] \lesssim HS\sqrt{\log(\cdot)} \cdot \sum_{k=1}^{K} \sum_{h=1}^{H} \mathbb{E}\left[\frac{1}{\sqrt{n_h^k(s_h^k, a_h^k)}}\right].$$

Basic Regret Bound

Theorem (Basic regret bound for UCB-VI)

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

$$\sum_{k=1}^{K} (V_0^\star - V_0^{\pi^k}) \;\le\; H^2 S \sqrt{8AK \log(2KHSA/\delta)}.$$

Proof sketch: On $\mathcal{E}_{\mathrm{opt}} \cap \mathcal{E}_{\mathrm{comp}}$, combine the certificate with the surplus bound:

$$\sum_{k=1}^{K} (V_0^\star - V_0^{\pi^k}) \le \sum_{k=1}^{K} \sum_{h=1}^{H} \mathbb{E}[\varepsilon_h^k(s_h^k, a_h^k)] \;\lesssim\; HS\sqrt{\log(\cdot)} \cdot \sum_{k=1}^{K} \sum_{h=1}^{H} \mathbb{E}\left[\frac{1}{\sqrt{n_h^k(s_h^k, a_h^k)}} \right].$$

Counting argument: For each fixed h, the j-th visit to any pair (s, a) contributes $j^{-1/2}$. Using $\sum_{j=1}^{m} j^{-1/2} \le 2\sqrt{m}$ and Cauchy–Schwarz:

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

Basic Regret Bound

Theorem (Basic regret bound for UCB-VI)

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

$$\sum_{k=1}^{K} (V_0^\star - V_0^{\pi^k}) \;\le\; H^2 S \sqrt{8AK \log(2KHSA/\delta)}.$$

Proof sketch: On $\mathcal{E}_{\mathrm{opt}} \cap \mathcal{E}_{\mathrm{comp}}$, combine the certificate with the surplus bound:

$$\sum_{k=1}^{K} (V_0^\star - V_0^{\pi^k}) \le \sum_{k=1}^{K} \sum_{h=1}^{H} \mathbb{E}[\varepsilon_h^k(s_h^k, a_h^k)] \;\lesssim\; HS\sqrt{\log(\cdot)} \cdot \sum_{k=1}^{K} \sum_{h=1}^{H} \mathbb{E}\left[\frac{1}{\sqrt{n_h^k(s_h^k, a_h^k)}}\right].$$

Counting argument: For each fixed h, the j-th visit to any pair (s, a) contributes $j^{-1/2}$. Using $\sum_{j=1}^{m} j^{-1/2} \le 2\sqrt{m}$ and Cauchy–Schwarz:

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

Summing over H stages gives total regret $\lesssim HS\sqrt{\log(\cdot)} \cdot H\sqrt{SAK} = H^2 S \sqrt{SAK \log(\cdot)}$.

□

Today's Plan

The $\sqrt{S}$ Bottleneck

The basic bound used $\mathcal{E}_{\text{comp}}$ to control $((\widehat{P}_h^k - P_h)\widehat{V}_{h+1}^k)$ *uniformly* over all V, incurring a $\sqrt{S}$ overhead:

$$\varepsilon_h^k(s,a) \lesssim H\sqrt{\frac{S\log(\cdot)}{n_h^k(s,a)}}.$$

The $\sqrt{S}$ Bottleneck

The basic bound used $\mathcal{E}_{\text{comp}}$ to control $((\widehat{P}_h^k - P_h)\widehat{V}_{h+1}^k)$ *uniformly* over all V, incurring a $\sqrt{S}$ overhead:

$$\varepsilon_h^k(s,a) \lesssim H\sqrt{\frac{S\log(\cdot)}{n_h^k(s,a)}}.$$

Key observation: we only need to control the gap $f = \widehat{V}_{h+1}^k - V_{h+1}^\star \geq 0$, not all of $\widehat{V}_{h+1}^k$. This gap can be small even when $\widehat{V}_{h+1}^k$ itself is large.

The $\sqrt{S}$ Bottleneck

The basic bound used $\mathcal{E}_{\text{comp}}$ to control $((\widehat{P}_h^k - P_h)\widehat{V}_{h+1}^k)$ *uniformly* over all V, incurring a $\sqrt{S}$ overhead:

$$\varepsilon_h^k(s,a) \lesssim H\sqrt{\frac{S\log(\cdot)}{n_h^k(s,a)}}.$$

Key observation: we only need to control the gap $f = \widehat{V}_{h+1}^k - V_{h+1}^\star \geq 0$, not all of $\widehat{V}_{h+1}^k$. This gap can be small even when $\widehat{V}_{h+1}^k$ itself is large.

Bernstein inequality for transitions

With prob. $\geq 1-\delta$, simultaneously for all (h,s,a,k) and all $f : \mathcal{S} \to [0,H]$:

$$\left|((P_h - \widehat{P}_h^k)f)(s,a)\right| \leq \frac{2H^2 S\,\iota_K}{n_h^k(s,a)} + \frac{1}{H}(P_h f)(s,a),$$

where $\iota_K = \log(2KHS^2A/\delta)$.

The $\sqrt{S}$ Bottleneck

The basic bound used $\mathcal{E}_{\text{comp}}$ to control $((\widehat{P}_h^k - P_h)\widehat{V}_{h+1}^k)$ *uniformly* over all V, incurring a $\sqrt{S}$ overhead:

$$\varepsilon_h^k(s,a) \lesssim H\sqrt{\frac{S\log(\cdot)}{n_h^k(s,a)}}.$$

Key observation: we only need to control the gap $f = \widehat{V}_{h+1}^k - V_{h+1}^\star \geq 0$, not all of $\widehat{V}_{h+1}^k$. This gap can be small even when $\widehat{V}_{h+1}^k$ itself is large.

Bernstein inequality for transitions

With prob. $\geq 1-\delta$, simultaneously for all (h,s,a,k) and all $f : \mathcal{S} \to [0,H]$:

$$\left|((P_h - \widehat{P}_h^k)f)(s,a)\right| \leq \frac{2H^2 S\,\iota_K}{n_h^k(s,a)} + \frac{1}{H}(P_h f)(s,a),$$

where $\iota_K = \log(2KHS^2A/\delta)$.

Key feature: the bound depends on $(P_h f)(s,a)$ — the expectation of f under the true transition — rather than $H\sqrt{S}$.

Proving the Bernstein Bound

Step 1 — Coordinatewise Bernstein. Fix (h, s, a, k), $n = n_h^k(s, a)$. For each $s' \in \mathcal{S}$, the indicators $\mathbf{1}[s_{h+1}^i = s']$ over the n visits are i.i.d. $\mathrm{Bernoulli}(P_h(s'|s, a))$, independent of any f. Bernstein's inequality gives:

$$\left|\widehat{P}_h^k(s'|s, a) - P_h(s'|s, a)\right| \;\le\; \sqrt{\frac{2P_h(s'|s, a)\,\iota_K}{n}} + \frac{2\iota_K}{3n},$$

where $\iota_K = \log(2KHS^2A/\delta)$. Union-bound over all (s', h, s, a, k) (at most KHS^2A tuples). By the triangle inequality, simultaneously for *all* $f : \mathcal{S} \to [0, H]$:

$$\left|((\widehat{P}_h^k - P_h)f)(s, a)\right| \;\le\; \sum_{s'} f(s')\left(\sqrt{\frac{2P_h(s')\,\iota_K}{n}} + \frac{2\iota_K}{3n}\right).$$

Proving the Bernstein Bound

Step 1 — Coordinatewise Bernstein. Fix (h, s, a, k), $n = n_h^k(s, a)$. For each $s' \in \mathcal{S}$, the indicators $\mathbf{1}[s_{h+1}^i = s']$ over the n visits are i.i.d. $\mathrm{Bernoulli}(P_h(s'|s, a))$, independent of any f. Bernstein's inequality gives:

$$\left|\widehat{P}_h^k(s'|s, a) - P_h(s'|s, a)\right| \;\le\; \sqrt{\frac{2P_h(s'|s, a)\,\iota_K}{n}} + \frac{2\iota_K}{3n},$$

where $\iota_K = \log(2KHS^2A/\delta)$. Union-bound over all (s', h, s, a, k) (at most KHS^2A tuples). By the triangle inequality, simultaneously for *all* $f : \mathcal{S} \to [0, H]$:

$$\left|((\widehat{P}_h^k - P_h)f)(s, a)\right| \;\le\; \sum_{s'} f(s')\left(\sqrt{\frac{2P_h(s')\,\iota_K}{n}} + \frac{2\iota_K}{3n}\right).$$

Step 2 — Cauchy–Schwarz.

$\sum_{s'} f(s')\sqrt{P_h(s')} \le \sqrt{\sum_{s'} P_h(s')f(s')} \cdot \sqrt{\sum_{s'} f(s')} \le \sqrt{(P_h f)(s, a)} \cdot \sqrt{HS}$, so Step 1 gives

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

Proving the Bernstein Bound

Step 1 — Coordinatewise Bernstein. Fix (h, s, a, k), $n = n_h^k(s, a)$. For each $s' \in \mathcal{S}$, the indicators $\mathbf{1}[s_{h+1}^i = s']$ over the n visits are i.i.d. $\text{Bernoulli}(P_h(s'|s, a))$, independent of any f. Bernstein's inequality gives:

$$\left|\widehat{P}_h^k(s'|s, a) - P_h(s'|s, a)\right| \;\le\; \sqrt{\frac{2P_h(s'|s, a)\,\iota_K}{n}} + \frac{2\iota_K}{3n},$$

where $\iota_K = \log(2KHS^2A/\delta)$. Union-bound over all (s', h, s, a, k) (at most KHS^2A tuples). By the triangle inequality, simultaneously for *all* $f : \mathcal{S} \to [0, H]$:

$$\left|((\widehat{P}_h^k - P_h)f)(s, a)\right| \;\le\; \sum_{s'} f(s')\left(\sqrt{\frac{2P_h(s')\,\iota_K}{n}} + \frac{2\iota_K}{3n}\right).$$

Step 2 — Cauchy–Schwarz.

$\sum_{s'} f(s')\sqrt{P_h(s')} \le \sqrt{\sum_{s'} P_h(s')f(s')} \cdot \sqrt{\sum_{s'} f(s')} \le \sqrt{(P_hf)(s, a)} \cdot \sqrt{HS}$, so Step 1 gives

$$\frac{2HS\,\iota_K}{n} + \sqrt{\frac{2HS\,\iota_K}{n} \cdot (P_hf)(s, a)}.$$

Step 3 — AM-GM. Apply $\sqrt{AB} \le \frac{1}{2}(\lambda A + B/\lambda)$ with $\lambda = H$, $A = 2HS\iota_K/n$, $B = (P_hf)$:

$$\sqrt{\frac{2HS\,\iota_K}{n} \cdot (P_hf)} \;\le\; \frac{H^2S\,\iota_K}{n} + \frac{(P_hf)}{2H}.$$

Combining (absorbing $\frac{2HS\iota_K}{n} \le \frac{H^2S\iota_K}{n}$ and $\frac{1}{2H} \le \frac{1}{H}$):

$$\left|((\widehat{P}_h^k - P_h)f)(s, a)\right| \;\le\; \frac{2H^2S\,\iota_K}{n_h^k(s, a)} + \frac{1}{H}(P_hf)(s, a). \qquad \square$$

Local Recursion for UCB-VI

Lemma (Local recursion)

Define $\Gamma_h^k := \widehat{V}_h^k - V_h^{\pi^k}$ and

$$\beta_h^k(s,a) := H\sqrt{\frac{2\log(2KHSA/\delta)}{n_h^k(s,a)}} + \frac{2H^2 S\,\iota_K}{n_h^k(s,a)}.$$

On the good event, for every k, h, s, with $a = \pi_h^k(s)$:

$$\varepsilon_h^k(s,a) \;\le\; \beta_h^k(s,a) + \frac{1}{H}(P_h\Gamma_{h+1}^k)(s,a).$$

Unrolling this recursion (using $(1+1/H)^H \le e$) yields:

$$V_0^\star - V_0^{\pi^k} \;\le\; e\sum_{h=1}^{H}\mathbb{E}\left[\beta_h^k(s_h^k, a_h^k)\right].$$

Local Recursion: Proof Sketch

Decompose the surplus by splitting $\widehat{V}^k_{h+1} = V^\star_{h+1} + (\widehat{V}^k_{h+1} - V^\star_{h+1})$:

$$\begin{aligned}
\varepsilon^k_h(s,a) &= \widehat{Q}^k_h(s,a) - r_h(s,a) - (P_h \widehat{V}^k_{h+1})(s,a) \\
&= b^k_h(s,a) + \big((\widehat{P}^k_h - P_h)\widehat{V}^k_{h+1}\big)(s,a) \\
&= b^k_h(s,a) + \big((\widehat{P}^k_h - P_h)V^\star_{h+1}\big)(s,a) + \big((\widehat{P}^k_h - P_h)(\widehat{V}^k_{h+1} - V^\star_{h+1})\big)(s,a).
\end{aligned}$$

Local Recursion: Proof Sketch

Decompose the surplus by splitting $\widehat{V}^k_{h+1} = V^\star_{h+1} + (\widehat{V}^k_{h+1} - V^\star_{h+1})$:

$$\begin{aligned}\varepsilon^k_h(s,a) &= \widehat{Q}^k_h(s,a) - r_h(s,a) - (P_h \widehat{V}^k_{h+1})(s,a) \\ &= b^k_h(s,a) + \big((\widehat{P}^k_h - P_h)\widehat{V}^k_{h+1}\big)(s,a) \\ &= b^k_h(s,a) + \big((\widehat{P}^k_h - P_h)V^\star_{h+1}\big)(s,a) + \big((\widehat{P}^k_h - P_h)(\widehat{V}^k_{h+1} - V^\star_{h+1})\big)(s,a).\end{aligned}$$

- **First two terms:** on $\mathcal{E}_{\mathrm{opt}}$, $|(\widehat{P}^k_h - P_h)V^\star_{h+1}| \le b^k_h$, so these sum to $\le 2b^k_h$.
- **Third term:** apply Bernstein with $f = \widehat{V}^k_{h+1} - V^\star_{h+1} \ge 0$:

$$\big((\widehat{P}^k_h - P_h)f\big)(s,a) \;\le\; \frac{2H^2 S\iota_K}{n^k_h(s,a)} + \frac{1}{H}(P_h f)(s,a) \;\le\; \frac{2H^2 S\iota_K}{n^k_h(s,a)} + \frac{1}{H}(P_h \Gamma^k_{h+1})(s,a),$$

since $f \le \Gamma^k_{h+1}$ because $V^\star_{h+1} \ge V^{\pi^k}_{h+1}$.

Local Recursion: Proof Sketch

Decompose the surplus by splitting $\widehat{V}^k_{h+1} = V^\star_{h+1} + (\widehat{V}^k_{h+1} - V^\star_{h+1})$:

$$\begin{aligned}\varepsilon^k_h(s,a) &= \widehat{Q}^k_h(s,a) - r_h(s,a) - (P_h \widehat{V}^k_{h+1})(s,a) \\ &= b^k_h(s,a) + \big((\widehat{P}^k_h - P_h)\widehat{V}^k_{h+1}\big)(s,a) \\ &= b^k_h(s,a) + \big((\widehat{P}^k_h - P_h)V^\star_{h+1}\big)(s,a) + \big((\widehat{P}^k_h - P_h)(\widehat{V}^k_{h+1} - V^\star_{h+1})\big)(s,a).\end{aligned}$$

- **First two terms:** on $\mathcal{E}_{\mathrm{opt}}$, $|(\widehat{P}^k_h - P_h)V^\star_{h+1}| \le b^k_h$, so these sum to $\le 2b^k_h$.
- **Third term:** apply Bernstein with $f = \widehat{V}^k_{h+1} - V^\star_{h+1} \ge 0$:

$$\big((\widehat{P}^k_h - P_h)f\big)(s,a) \;\le\; \frac{2H^2 S\iota_K}{n^k_h(s,a)} + \frac{1}{H}(P_h f)(s,a) \;\le\; \frac{2H^2 S\iota_K}{n^k_h(s,a)} + \frac{1}{H}(P_h \Gamma^k_{h+1})(s,a),$$

since $f \le \Gamma^k_{h+1}$ because $V^\star_{h+1} \ge V^{\pi^k}_{h+1}$.

Combining: $\varepsilon^k_h \le \beta^k_h + \frac{1}{H}(P_h\Gamma^k_{h+1})$. Since $\Gamma^k_h = \varepsilon^k_h + (P_h\Gamma^k_{h+1})$, the recursion gives $\Gamma^k_h \le \beta^k_h + (1+\frac{1}{H})(P_h\Gamma^k_{h+1})$. Unrolling and using $(1+1/H)^H \le e$ yields the certificate.

Improved Regret Bound

Theorem (Improved regret bound)

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

$$\sum_{k=1}^{K}(V_0^\star - V_0^{\pi^k}) \leq H^2\sqrt{8e^2SAK\log(2KHSA/\delta)} + 4eH^3S^2A\log(K)\log(2KHS^2A/\delta).$$

Improved Regret Bound

Theorem (Improved regret bound)

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

$$\sum_{k=1}^{K}(V_0^\star - V_0^{\pi^k}) \leq H^2\sqrt{8e^2SAK\log(2KHSA/\delta)} + 4eH^3S^2A\log(K)\log(2KHS^2A/\delta).$$

Proof: Sum the local certificate over k. Two separate counting arguments:

- $n^{-1/2}$ **term:** $\sum_{k,h}(n_h^k)^{-1/2} \leq 2H\sqrt{SAK}$ *(same counting as basic proof).*
- n^{-1} **term:** $\sum_{k,h}(n_h^k)^{-1} = \sum_{h,s,a}\sum_{j=1}^{n_h^{K+1}(s,a)} j^{-1} \leq 2SAH\log K.$

Improved Regret Bound

Theorem (Improved regret bound)

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

$$\sum_{k=1}^{K}(V_0^\star - V_0^{\pi^k}) \;\le\; H^2\sqrt{8e^2 SAK \log(2KHSA/\delta)} + 4eH^3S^2A\log(K)\log(2KHS^2A/\delta).$$

Proof: Sum the local certificate over k. Two separate counting arguments:

- $n^{-1/2}$ **term:** $\sum_{k,h}(n_h^k)^{-1/2} \le 2H\sqrt{SAK}$ *(same counting as basic proof).*
- n^{-1} **term:** $\sum_{k,h}(n_h^k)^{-1} = \sum_{h,s,a}\sum_{j=1}^{n_h^{K+1}(s,a)} j^{-1} \le 2SAH\log K.$

Summary of regret bounds

Basic bound (via $\mathcal{E}_{\mathrm{comp}}$)	$\widetilde{O}(H^2S\sqrt{SAK})$
Improved bound (Bernstein)	$\widetilde{O}(H^2\sqrt{SAK} + H^3S^2A)$
Minimax optimal	$\widetilde{\Theta}(H^{3/2}\sqrt{SAK})$

Today's Plan

Gap-Dependent Bounds

The minimax bound treats every episode as equally hard.

Intuition: if action a at (s, h) is clearly suboptimal, UCB-VI should stop exploring it after a relatively small number of visits — much fewer than the $\sqrt{K}$ rate suggests.

Gap-Dependent Bounds

The minimax bound treats every episode as equally hard.

Intuition: if action a at (s, h) is clearly suboptimal, UCB-VI should stop exploring it after a relatively small number of visits — much fewer than the $\sqrt{K}$ rate suggests.

Sub-optimality gap

$$\Delta_h(s, a) := V_h^\star(s) - Q_h^\star(s, a) \geq 0.$$

$\Delta_h(s, a) = 0$ iff a is an optimal action at (s, h).

Gap-Dependent Bounds

The minimax bound treats every episode as equally hard.

Intuition: if action a at (s, h) is clearly suboptimal, UCB-VI should stop exploring it after a relatively small number of visits — much fewer than the $\sqrt{K}$ rate suggests.

Sub-optimality gap

$$\Delta_h(s, a) := V_h^\star(s) - Q_h^\star(s, a) \;\geq\; 0.$$

$\Delta_h(s, a) = 0$ iff a is an optimal action at (s, h).

Goal: show that the total regret is bounded by a *logarithmic* sum over inverse gaps:

$$\sum_{k=1}^{K} \left(V_0^\star - V_0^{\pi^k}\right) \;\lesssim\; \sum_{h=1}^{H} \sum_{\substack{s,a: \\ \Delta_h(s,a)>0}} \frac{H^3\, \iota_K}{\Delta_h(s, a)}.$$

This is $O(\log K)$ whenever all gaps are bounded away from zero.

Performance Difference Lemma

Lemma (Performance difference)

For any two policies π and ρ:

$$V_0^\pi - V_0^\rho = \sum_{h=1}^{H} \sum_{s,a} w_h^\pi(s,a) \left(Q_h^\rho(s,a) - V_h^\rho(s) \right).$$

Taking $\rho = \pi^\star$ gives:

$$V_0^\star - V_0^\pi = \sum_{h=1}^{H} \sum_{s,a} w_h^\pi(s,a) \, \Delta_h(s,a).$$

Performance Difference Lemma

Lemma (Performance difference)

For any two policies π and ρ:

$$V_0^\pi - V_0^\rho = \sum_{h=1}^{H} \sum_{s,a} w_h^\pi(s,a) \left(Q_h^\rho(s,a) - V_h^\rho(s)\right).$$

Taking $\rho = \pi^\star$ gives:

$$V_0^\star - V_0^\pi = \sum_{h=1}^{H} \sum_{s,a} w_h^\pi(s,a)\, \Delta_h(s,a).$$

Proof sketch: For each (h,s), expand $V_h^\pi - V_h^\rho$ using both policies' Bellman equations:

$$V_h^\pi(s) - V_h^\rho(s) = \sum_a \pi_h(a|s)\big(Q_h^\rho(s,a) - V_h^\rho(s)\big) + (P_h(V_{h+1}^\pi - V_{h+1}^\rho))(s,\pi_h(s)).$$

Unroll from $h=1$ to H and average over $s_1 \sim \nu$. □

Performance Difference Lemma

Lemma (Performance difference)

For any two policies π and ρ:

$$V_0^\pi - V_0^\rho = \sum_{h=1}^{H} \sum_{s,a} w_h^\pi(s,a) \left(Q_h^\rho(s,a) - V_h^\rho(s)\right).$$

Taking $\rho = \pi^\star$ gives:

$$V_0^\star - V_0^\pi = \sum_{h=1}^{H} \sum_{s,a} w_h^\pi(s,a)\, \Delta_h(s,a).$$

Proof sketch: For each (h,s), expand $V_h^\pi - V_h^\rho$ using both policies' Bellman equations:

$$V_h^\pi(s) - V_h^\rho(s) = \sum_a \pi_h(a|s)\left(Q_h^\rho(s,a) - V_h^\rho(s)\right) + (P_h(V_{h+1}^\pi - V_{h+1}^\rho))(s,\pi_h(s)).$$

Unroll from $h = 1$ to H and average over $s_1 \sim \nu$. □

Consequence

The Clipping Idea: Setup

Fix episode k, stage h, state s, and let $a = \pi_h^k(s)$.

Fact 1 (local recursion):

$$\Gamma_h^k(s) \;\le\; \beta_h^k(s,a) + D_h^k(s,a), \qquad D_h^k(s,a) := \left(1 + \tfrac{1}{H}\right)(P_h \Gamma_{h+1}^k)(s,a). \qquad \text{(LR)}$$

The Clipping Idea: Setup

Fix episode k, stage h, state s, and let $a = \pi_h^k(s)$.

Fact 1 (local recursion):

$$\Gamma_h^k(s) \;\le\; \beta_h^k(s,a) + D_h^k(s,a), \qquad D_h^k(s,a) := \left(1 + \tfrac{1}{H}\right)(P_h \Gamma_{h+1}^k)(s,a). \tag{LR}$$

Fact 2: the *same* bound holds for the gap $\Delta_h(s,a) = V_h^\star(s) - Q_h^\star(s,a)$:

$$\Delta_h(s,a) \;\le\; \beta_h^k(s,a) + D_h^k(s,a). \tag{$\star$}$$

The Clipping Idea: Setup

Fix episode k, stage h, state s, and let $a = \pi_h^k(s)$.

Fact 1 (local recursion):

$$\Gamma_h^k(s) \;\le\; \beta_h^k(s,a) + D_h^k(s,a), \qquad D_h^k(s,a) := \left(1 + \tfrac{1}{H}\right)(P_h \Gamma_{h+1}^k)(s,a). \tag{LR}$$

Fact 2: the *same* bound holds for the gap $\Delta_h(s,a) = V_h^\star(s) - Q_h^\star(s,a)$:

$$\Delta_h(s,a) \;\le\; \beta_h^k(s,a) + D_h^k(s,a). \tag{$\star$}$$

Step 1. Since UCB-VI is greedy, $\widehat{Q}_h^k(s,a) = \widehat{V}_h^k(s)$. By optimism $\widehat{V}_h^k(s) \ge V_h^\star(s)$, so subtracting $Q_h^\star(s,a)$ from both sides:

$$\Delta_h(s,a) = V_h^\star(s) - Q_h^\star(s,a) \;\le\; \widehat{Q}_h^k(s,a) - Q_h^\star(s,a).$$

The Clipping Idea: Setup

Fix episode k, stage h, state s, and let $a = \pi_h^k(s)$.

Fact 1 (local recursion):

$$\Gamma_h^k(s) \;\le\; \beta_h^k(s,a) + D_h^k(s,a), \qquad D_h^k(s,a) := \left(1 + \tfrac{1}{H}\right)(P_h \Gamma_{h+1}^k)(s,a). \qquad \text{(LR)}$$

Fact 2: the *same* bound holds for the gap $\Delta_h(s,a) = V_h^\star(s) - Q_h^\star(s,a)$:

$$\Delta_h(s,a) \;\le\; \beta_h^k(s,a) + D_h^k(s,a). \qquad (\star)$$

Step 1. Since UCB-VI is greedy, $\widehat{Q}_h^k(s,a) = \widehat{V}_h^k(s)$. By optimism $\widehat{V}_h^k(s) \ge V_h^\star(s)$, so subtracting $Q_h^\star(s,a)$ from both sides:

$$\Delta_h(s,a) = V_h^\star(s) - Q_h^\star(s,a) \;\le\; \widehat{Q}_h^k(s,a) - Q_h^\star(s,a).$$

Step 2. It remains to show $\widehat{Q}_h^k(s,a) - Q_h^\star(s,a) \le \beta_h^k(s,a) + D_h^k(s,a)$.
See next slide.

The Clipping Idea: Proving Fact 2

Step 2 (continued). Decompose $\widehat{Q}_h^k(s,a) - Q_h^\star(s,a)$.
Using $Q_h^\star = r_h + (P_h V_{h+1}^\star)$ and $\widehat{Q}_h^k = r_h + (\widehat{P}_h^k \widehat{V}_{h+1}^k) + b_h^k$:

$$\widehat{Q}_h^k(s,a) - Q_h^\star(s,a) = \underbrace{\left[\widehat{Q}_h^k - r_h - (P_h \widehat{V}_{h+1}^k)\right](s,a)}_{= \varepsilon_h^k(s,a)} + \left(P_h(\widehat{V}_{h+1}^k - V_{h+1}^\star)\right)(s,a).$$

The Clipping Idea: Proving Fact 2

Step 2 (continued). Decompose $\widehat{Q}_h^k(s,a) - Q_h^\star(s,a)$.
Using $Q_h^\star = r_h + (P_h V_{h+1}^\star)$ and $\widehat{Q}_h^k = r_h + (\widehat{P}_h^k \widehat{V}_{h+1}^k) + b_h^k$:

$$\widehat{Q}_h^k(s,a) - Q_h^\star(s,a) = \underbrace{\left[\widehat{Q}_h^k - r_h - (P_h \widehat{V}_{h+1}^k)\right](s,a)}_{= \varepsilon_h^k(s,a)} + \left(P_h(\widehat{V}_{h+1}^k - V_{h+1}^\star)\right)(s,a).$$

Bound the two pieces separately:

Piece 1: by the local recursion,

$$\varepsilon_h^k(s,a) \leq \beta_h^k(s,a) + \tfrac{1}{H}(P_h \Gamma_{h+1}^k)(s,a).$$

The Clipping Idea: Proving Fact 2

Step 2 (continued). Decompose $\widehat{Q}_h^k(s,a) - Q_h^\star(s,a)$.
Using $Q_h^\star = r_h + (P_h V_{h+1}^\star)$ and $\widehat{Q}_h^k = r_h + (\widehat{P}_h^k \widehat{V}_{h+1}^k) + b_h^k$:

$$\widehat{Q}_h^k(s,a) - Q_h^\star(s,a) = \underbrace{\big[\widehat{Q}_h^k - r_h - (P_h \widehat{V}_{h+1}^k)\big](s,a)}_{= \varepsilon_h^k(s,a)} + \big(P_h(\widehat{V}_{h+1}^k - V_{h+1}^\star)\big)(s,a).$$

Bound the two pieces separately:

Piece 1: by the local recursion,

$$\varepsilon_h^k(s,a) \;\le\; \beta_h^k(s,a) + \tfrac{1}{H}(P_h \Gamma_{h+1}^k)(s,a).$$

Piece 2: since $V_{h+1}^\star \ge V_{h+1}^{\pi^k}$, we have $\widehat{V}_{h+1}^k - V_{h+1}^\star \le \widehat{V}_{h+1}^k - V_{h+1}^{\pi^k} = \Gamma_{h+1}^k$, so

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

The Clipping Idea: Proving Fact 2

Step 2 (continued). Decompose $\widehat{Q}_h^k(s,a) - Q_h^\star(s,a)$.
Using $Q_h^\star = r_h + (P_h V_{h+1}^\star)$ and $\widehat{Q}_h^k = r_h + (\widehat{P}_h^k \widehat{V}_{h+1}^k) + b_h^k$:

$$\widehat{Q}_h^k(s,a) - Q_h^\star(s,a) = \underbrace{\left[\widehat{Q}_h^k - r_h - (P_h \widehat{V}_{h+1}^k)\right](s,a)}_{= \varepsilon_h^k(s,a)} + \left(P_h(\widehat{V}_{h+1}^k - V_{h+1}^\star)\right)(s,a).$$

Bound the two pieces separately:

Piece 1: by the local recursion,

$$\varepsilon_h^k(s,a) \le \beta_h^k(s,a) + \tfrac{1}{H}(P_h \Gamma_{h+1}^k)(s,a).$$

Piece 2: since $V_{h+1}^\star \ge V_{h+1}^{\pi^k}$, we have $\widehat{V}_{h+1}^k - V_{h+1}^\star \le \widehat{V}_{h+1}^k - V_{h+1}^{\pi^k} = \Gamma_{h+1}^k$, so

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

Adding Pieces 1 and 2:

$$\Delta_h(s,a) \le \beta_h^k(s,a) + \left(1 + \tfrac{1}{H}\right)(P_h \Gamma_{h+1}^k)(s,a) = \beta_h^k(s,a) + D_h^k(s,a). \quad \square$$

The Clipping Idea: Both Facts Together

Fix episode k, stage h, state s, and $a = \pi_h^k(s)$. We have established:

Fact 1 (local recursion):

$$\Gamma_h^k(s) \;\le\; \beta_h^k(s,a) + D_h^k(s,a), \qquad D_h^k(s,a) := \left(1 + \tfrac{1}{H}\right)(P_h \Gamma_{h+1}^k)(s,a). \qquad \text{(LR)}$$

Fact 2 (just proved):

$$\Delta_h(s,a) \;\le\; \beta_h^k(s,a) + D_h^k(s,a). \qquad (\star)$$

The Clipping Idea: Both Facts Together

Fix episode k, stage h, state s, and $a = \pi_h^k(s)$. We have established:

Fact 1 (local recursion):

$$\Gamma_h^k(s) \;\le\; \beta_h^k(s,a) + D_h^k(s,a), \qquad D_h^k(s,a) := \left(1 + \tfrac{1}{H}\right)(P_h \Gamma_{h+1}^k)(s,a). \qquad \text{(LR)}$$

Fact 2 (just proved):

$$\Delta_h(s,a) \;\le\; \beta_h^k(s,a) + D_h^k(s,a). \qquad (\star)$$

Both facts have the *same* right-hand side. The key question is whether $\beta_h^k(s,a)$ is large or small relative to the gap.

Define the **clip threshold**

$$\tau_h(s,a) \;:=\; \frac{\Delta_h(s,a)}{4H}.$$

- **Case 1** ($\beta_h^k \ge \tau_h$): local uncertainty is still "large" — exploration of this pair is not yet finished.
- **Case 2** ($\beta_h^k < \tau_h$): local uncertainty is "small" — but then $(\star)$ forces D_h^k to be large.

The Clipping Idea: Case Analysis

Case 1: $\beta_h^k(s,a) \geq \tau_h(s,a)$.
From (LR) directly:

$$\Gamma_h^k(s) \;\leq\; \beta_h^k(s,a) + D_h^k(s,a) \;=\; \underbrace{\beta_h^k(s,a)}_{\geq \tau_h,\ \text{not clipped}} + D_h^k(s,a).$$

The Clipping Idea: Case Analysis

Case 1: $\beta_h^k(s,a) \geq \tau_h(s,a)$.
From (LR) directly:

$$\Gamma_h^k(s) \;\leq\; \beta_h^k(s,a) + D_h^k(s,a) \;=\; \underbrace{\beta_h^k(s,a)}_{\geq \tau_h,\ \text{not clipped}} + D_h^k(s,a).$$

Case 2: $\beta_h^k(s,a) < \tau_h(s,a) = \Delta_h(s,a)/(4H)$.
From $(\star)$: $\Delta_h(s,a) \leq \beta_h^k(s,a) + D_h^k(s,a)$, so

$$D_h^k(s,a) \;\geq\; \Delta_h(s,a) - \beta_h^k(s,a) \;>\; \Delta_h(s,a) - \frac{\Delta_h(s,a)}{4H} = \frac{4H-1}{4H}\,\Delta_h(s,a).$$

The Clipping Idea: Case Analysis

Case 1: $\beta_h^k(s,a) \geq \tau_h(s,a)$.
From (LR) directly:

$$\Gamma_h^k(s) \;\leq\; \beta_h^k(s,a) + D_h^k(s,a) \;=\; \underbrace{\beta_h^k(s,a)}_{\geq \tau_h,\ \text{not clipped}} + D_h^k(s,a).$$

Case 2: $\beta_h^k(s,a) < \tau_h(s,a) = \Delta_h(s,a)/(4H)$.
From $(\star)$: $\Delta_h(s,a) \leq \beta_h^k(s,a) + D_h^k(s,a)$, so

$$D_h^k(s,a) \;\geq\; \Delta_h(s,a) - \beta_h^k(s,a) \;>\; \Delta_h(s,a) - \frac{\Delta_h(s,a)}{4H} = \frac{4H-1}{4H}\,\Delta_h(s,a).$$

In particular $\beta_h^k(s,a) < \tau_h = \Delta/(4H) < D_h^k/(4H-1)$. Substituting into (LR):

$$\Gamma_h^k(s) \;\leq\; \beta_h^k + D_h^k \;<\; \frac{D_h^k}{4H-1} + D_h^k = \left(1 + \tfrac{1}{4H-1}\right) D_h^k(s,a).$$

The Clipping Idea: Case Analysis

Case 1: $\beta_h^k(s,a) \geq \tau_h(s,a)$.
From (LR) directly:

$$\Gamma_h^k(s) \;\leq\; \beta_h^k(s,a) + D_h^k(s,a) \;=\; \underbrace{\beta_h^k(s,a)}_{\geq \tau_h,\ \text{not clipped}} + D_h^k(s,a).$$

Case 2: $\beta_h^k(s,a) < \tau_h(s,a) = \Delta_h(s,a)/(4H)$.
From $(\star)$: $\Delta_h(s,a) \leq \beta_h^k(s,a) + D_h^k(s,a)$, so

$$D_h^k(s,a) \;\geq\; \Delta_h(s,a) - \beta_h^k(s,a) \;>\; \Delta_h(s,a) - \frac{\Delta_h(s,a)}{4H} = \frac{4H-1}{4H}\,\Delta_h(s,a).$$

In particular $\beta_h^k(s,a) < \tau_h = \Delta/(4H) < D_h^k/(4H-1)$. Substituting into (LR):

$$\Gamma_h^k(s) \;\leq\; \beta_h^k + D_h^k \;<\; \frac{D_h^k}{4H-1} + D_h^k = \left(1 + \tfrac{1}{4H-1}\right) D_h^k(s,a).$$

Combining both cases (noting $\mathrm{clip}(\beta_h^k \mid \tau_h) = 0$ in Case 2):

$$\Gamma_h^k(s) \;\leq\; \underbrace{\mathrm{clip}\big(\beta_h^k(s,a) \mid \tau_h(s,a)\big)}_{\text{zero in Case 2}} + \left(1 + \tfrac{1}{4H-1}\right) D_h^k(s,a).$$

Unrolling from $h=1$ to H (using $((1+\frac{1}{4H-1})(1+\frac{1}{H}))^H \leq e^2$):

$$V_0^\star - V_0^{\pi^k} \;\leq\; e^2 \sum_{h,s,a} w_h^{\pi^k}(s,a) \cdot \beta_h^k(s,a) \cdot \mathbf{1}\Big\{\beta_h^k(s,a) \geq \tfrac{\Delta_h(s,a)}{4H}\Big\}.$$

Gap-Dependent Regret Bound

Counting argument: For each (h, s, a) with $\Delta_h(s, a) > 0$, the condition $\beta_h^k \geq \frac{\Delta_h(s,a)}{4H}$ can hold for at most

$$N_h(s, a) \lesssim \frac{H^4 \iota_K}{\Delta_h(s, a)^2}$$

total visits (since $\beta_h^k \sim H\sqrt{\iota_K / n}$ eventually falls below $\Delta_h(s, a)/(4H)$).

Gap-Dependent Regret Bound

Counting argument: For each (h, s, a) with $\Delta_h(s, a) > 0$, the condition $\beta_h^k \geq \frac{\Delta_h(s,a)}{4H}$ can hold for at most

$$N_h(s, a) \lesssim \frac{H^4 \iota_K}{\Delta_h(s, a)^2}$$

total visits (since $\beta_h^k \sim H\sqrt{\iota_K/n}$ eventually falls below $\Delta_h(s, a)/(4H)$).
Summing the dominant $H\sqrt{\iota_K/n}$ contribution over those first $N_h(s, a)$ visits:

$$\sum_{n=1}^{N_h(s,a)} H\sqrt{\frac{\iota_K}{n}} \lesssim H\sqrt{\iota_K} \cdot \sqrt{N_h(s, a)} \lesssim \frac{H^3 \iota_K}{\Delta_h(s, a)}.$$

Gap-Dependent Regret Bound

Counting argument: For each (h, s, a) with $\Delta_h(s, a) > 0$, the condition $\beta_h^k \geq \frac{\Delta_h(s,a)}{4H}$ can hold for at most

$$N_h(s, a) \lesssim \frac{H^4 \iota_K}{\Delta_h(s, a)^2}$$

total visits (since $\beta_h^k \sim H\sqrt{\iota_K/n}$ eventually falls below $\Delta_h(s, a)/(4H)$).
Summing the dominant $H\sqrt{\iota_K/n}$ contribution over those first $N_h(s, a)$ visits:

$$\sum_{n=1}^{N_h(s,a)} H\sqrt{\frac{\iota_K}{n}} \lesssim H\sqrt{\iota_K} \cdot \sqrt{N_h(s, a)} \lesssim \frac{H^3 \iota_K}{\Delta_h(s, a)}.$$

Theorem (Gap-dependent regret bound, informal)

With high probability:

$$\sum_{k=1}^{K}(V_0^\star - V_0^{\pi^k}) \lesssim \sum_{h=1}^{H} \sum_{\substack{s,a: \\ \Delta_h(s,a)>0}} \frac{H^3 \iota_K}{\Delta_h(s, a)} + \text{Zero-gap-terms}.$$

Gap-Dependent Regret Bound

Counting argument: For each (h, s, a) with $\Delta_h(s, a) > 0$, the condition $\beta_h^k \geq \frac{\Delta_h(s,a)}{4H}$ can hold for at most

$$N_h(s, a) \lesssim \frac{H^4 \iota_K}{\Delta_h(s, a)^2}$$

total visits (since $\beta_h^k \sim H\sqrt{\iota_K/n}$ eventually falls below $\Delta_h(s, a)/(4H)$).
Summing the dominant $H\sqrt{\iota_K/n}$ contribution over those first $N_h(s, a)$ visits:

$$\sum_{n=1}^{N_h(s,a)} H\sqrt{\frac{\iota_K}{n}} \lesssim H\sqrt{\iota_K} \cdot \sqrt{N_h(s, a)} \lesssim \frac{H^3 \iota_K}{\Delta_h(s, a)}.$$

Theorem (Gap-dependent regret bound, informal)

With high probability:

$$\sum_{k=1}^{K}(V_0^\star - V_0^{\pi^k}) \lesssim \sum_{h=1}^{H} \sum_{\substack{s,a: \\ \Delta_h(s,a)>0}} \frac{H^3 \iota_K}{\Delta_h(s, a)} + \text{Zero-gap-terms}.$$

Structure: the gap-dependent proof is the *same* local recursion as the minimax proof, with β_h^k replaced by its clipped version. No new algorithmic ideas are needed.

Gap-Dependent Regret Bound

Counting argument: For each (h, s, a) with $\Delta_h(s, a) > 0$, the condition $\beta_h^k \geq \frac{\Delta_h(s,a)}{4H}$ can hold for at most

$$N_h(s, a) \lesssim \frac{H^4 \iota_K}{\Delta_h(s, a)^2}$$

total visits (since $\beta_h^k \sim H\sqrt{\iota_K / n}$ eventually falls below $\Delta_h(s, a)/(4H)$).
Summing the dominant $H\sqrt{\iota_K / n}$ contribution over those first $N_h(s, a)$ visits:

$$\sum_{n=1}^{N_h(s,a)} H\sqrt{\frac{\iota_K}{n}} \lesssim H\sqrt{\iota_K} \cdot \sqrt{N_h(s, a)} \lesssim \frac{H^3 \iota_K}{\Delta_h(s, a)}.$$

Theorem (Gap-dependent regret bound, informal)

With high probability:

$$\sum_{k=1}^{K}(V_0^\star - V_0^{\pi^k}) \lesssim \sum_{h=1}^{H} \sum_{\substack{s,a: \\ \Delta_h(s,a)>0}} \frac{H^3 \iota_K}{\Delta_h(s, a)} + \text{Zero-gap-terms}.$$

Structure: the gap-dependent proof is the *same* local recursion as the minimax proof, with β_h^k replaced by its clipped version. No new algorithmic ideas are needed.

Remark. The bound above handles only pairs with $\Delta_h(s, a) > 0$. Optimal actions ($\Delta_h = 0$) require separate bookkeeping (they are never clipped. See Simchowitz & Jamieson (2019) for details.

Today's Plan

Why Regret Is the Wrong Metric for PAC

Recall online-to-batch: $\widetilde{O}(\sqrt{K})$ regret $\Rightarrow$ $\widetilde{O}(1/\sqrt{K})$ suboptimality.

Gap-dependent regret $\widetilde{O}(H^3/\Delta) \Rightarrow$ PAC complexity $\widetilde{O}(H^3/(\Delta\varepsilon))$ via online-to-batch.

Why Regret Is the Wrong Metric for PAC

Recall online-to-batch: $\widetilde{O}(\sqrt{K})$ regret $\Rightarrow$ $\widetilde{O}(1/\sqrt{K})$ suboptimality.

Gap-dependent regret $\widetilde{O}(H^3/\Delta) \Rightarrow$ PAC complexity $\widetilde{O}(H^3/(\Delta\varepsilon))$ via online-to-batch.

The fundamental gap

A low-regret algorithm must play *reward-maximizing* policies throughout learning.

A PAC algorithm is free to play *highly suboptimal* policies if they reach informative states faster.

Why Regret Is the Wrong Metric for PAC

Recall online-to-batch: $\widetilde{O}(\sqrt{K})$ regret $\Rightarrow$ $\widetilde{O}(1/\sqrt{K})$ suboptimality.

Gap-dependent regret $\widetilde{O}(H^3/\Delta) \Rightarrow$ PAC complexity $\widetilde{O}(H^3/(\Delta\varepsilon))$ via online-to-batch.

The fundamental gap

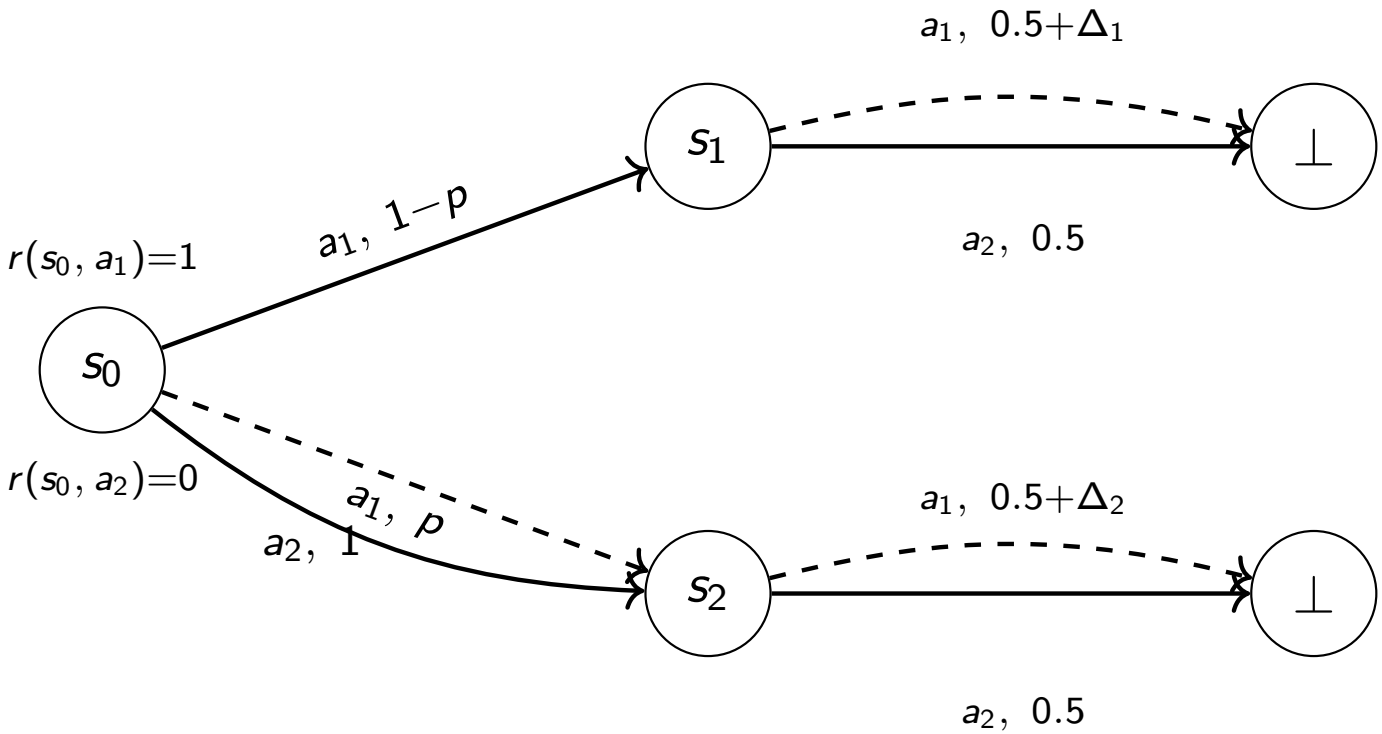
A low-regret algorithm must play *reward-maximizing* policies throughout learning.

A PAC algorithm is free to play *highly suboptimal* policies if they reach informative states faster.

Example (Wagenmaker et al., 2022):

- At s_0: action a_1 is optimal (reward 1) but reaches hard state s_2 with prob. $p \ll 1$.
- Action a_2 is suboptimal at s_0 (reward 0) but reaches s_2 with prob. 1.
- A low-regret algorithm mostly plays a_1: visits s_2 only $\approx pK$ times.
- A PAC algorithm plays a_2 during exploration: visits s_2 every episode.
- **Speedup factor:** $1/p$ in the cost to identify the best action at s_2.

The Motivating MDP



Low-regret algorithm (mostly plays a_1):

$$\frac{\log(1/\delta)}{\Delta_1^2} + \frac{\log(1/\delta)}{p\,\Delta_2^2}$$

PAC-focused algorithm (uses a_2 to explore):

$$\frac{\log(1/\delta)}{\Delta_1^2} + \frac{\log(1/\delta)}{\Delta_2^2}$$

A factor of $1/p$ improvement in the second term.

The Instance-Dependent PAC Complexity

Exact-identification complexity $C^\star(M)$ (Wagenmaker et al., 2022)

$$C^\star(M) := \sum_{h=1}^{H} \inf_{\pi} \max_{s \in \mathcal{S}} \frac{1}{w_h^\pi(s)} \sum_{a:\, \Delta_h(s,a)>0} \frac{1}{\Delta_h(s,a)^2}.$$

The Instance-Dependent PAC Complexity

Exact-identification complexity $C^\star(M)$ (Wagenmaker et al., 2022)

$$C^\star(M) := \sum_{h=1}^{H} \inf_{\pi} \max_{s \in \mathcal{S}} \frac{1}{w_h^\pi(s)} \sum_{a:\, \Delta_h(s,a)>0} \frac{1}{\Delta_h(s,a)^2}.$$

Two forces in tension:

- **Reachability:** $w_h^\pi(s)$ — how often roll-in policy π visits state s at stage h. We want this large for all hard states simultaneously.
- **Local identification:** $\sum_{a:\Delta_h>0} \Delta_h(s,a)^{-2}$ — the cost to identify $\pi_h^\star(s)$, as in best-arm identification.

The Instance-Dependent PAC Complexity

Exact-identification complexity $C^\star(M)$ (Wagenmaker et al., 2022)

$$C^\star(M) := \sum_{h=1}^{H} \inf_{\pi} \max_{s \in \mathcal{S}} \frac{1}{w_h^\pi(s)} \sum_{a:\, \Delta_h(s,a)>0} \frac{1}{\Delta_h(s,a)^2}.$$

Two forces in tension:

- **Reachability:** $w_h^\pi(s)$ — how often roll-in policy π visits state s at stage h. We want this large for all hard states simultaneously.
- **Local identification:** $\sum_{a:\Delta_h>0} \Delta_h(s,a)^{-2}$ — the cost to identify $\pi_h^\star(s)$, as in best-arm identification.

Optimal within-state allocation: Given $w_h^\pi(s)$ visits to state s, split them across actions as $\lambda_h(a|s) \propto \Delta_h(s,a)^{-2}$. This minimizes the max-over-a sample count, giving the sum $\sum_a \Delta_h(s,a)^{-2}$ as the per-state cost.

Where $C^\star(M)$ Comes From

Backward induction on stages.

Where $C^\star(M)$ Comes From

Backward induction on stages.

Stage $h = H$: Given arrival at state s, identifying $\pi^\star_H(s)$ is *exactly* a best-arm identification problem. Action a needs $\approx \log(1/\delta)/\Delta_H(s,a)^2$ samples.

Where $C^\star(M)$ Comes From

Backward induction on stages.

Stage $h = H$: Given arrival at state s, identifying $\pi^\star_H(s)$ is *exactly* a best-arm identification problem. Action a needs $\approx \log(1/\delta)/\Delta_H(s,a)^2$ samples.

Inductive step: Suppose $\pi^\star_{h+1:H}$ is already known. Starting at (s,a,h) and rolling out $\pi^\star_{h+1:H}$ yields a sample with mean $Q^\star_h(s,a)$. Conditional on reaching s at stage h: *again a best-arm identification problem* with gaps $\{\Delta_h(s,a)\}$.

Where $C^\star(M)$ Comes From

Backward induction on stages.

Stage $h = H$: Given arrival at state s, identifying $\pi^\star_H(s)$ is *exactly* a best-arm identification problem. Action a needs $\approx \log(1/\delta)/\Delta_H(s,a)^2$ samples.

Inductive step: Suppose $\pi^\star_{h+1:H}$ is already known. Starting at (s,a,h) and rolling out $\pi^\star_{h+1:H}$ yields a sample with mean $Q^\star_h(s,a)$. Conditional on reaching s at stage h: *again a best-arm identification problem* with gaps $\{\Delta_h(s,a)\}$.

Reachability constraint: With roll-in π and within-state allocation $\lambda_h(\cdot|s)$, action a at (s,h) receives $\approx K\, w^\pi_h(s)\, \lambda_h(a|s)$ samples. Requiring these to exceed $\log(1/\delta)/\Delta_h(s,a)^2$ gives:

$$K \gtrsim \frac{\log(1/\delta)}{w^\pi_h(s)} \max_a \frac{1}{\lambda_h(a|s)\,\Delta_h(s,a)^2}.$$

Optimize λ and take the maximum over states to get the stage-h cost. Sum over h. □

Takeaways

Key messages from this lecture

1. **UCB-VI** adds optimistic bonuses to Bellman backups: basic regret bound $\widetilde{O}(H^2 S\sqrt{SAK})$.
2. **Bernstein inequality + local recursion** removes the $\sqrt{S}$ overhead: improved bound $\widetilde{O}(H^2\sqrt{SAK})$.
3. **Gap-dependent regret** $\widetilde{O}(H^3/\Delta)$ via the same local recursion with a clip — no new algorithm needed.
4. **PAC $\neq$ low regret**: deliberately exploring via suboptimal actions can be faster by a factor $1/p$ when the interesting state is hard to reach.
5. **Instance-optimal PAC complexity** is $C^\star(M)$: a stage-wise sum of inverse-gap-squared terms, weighted by the reachability cost.