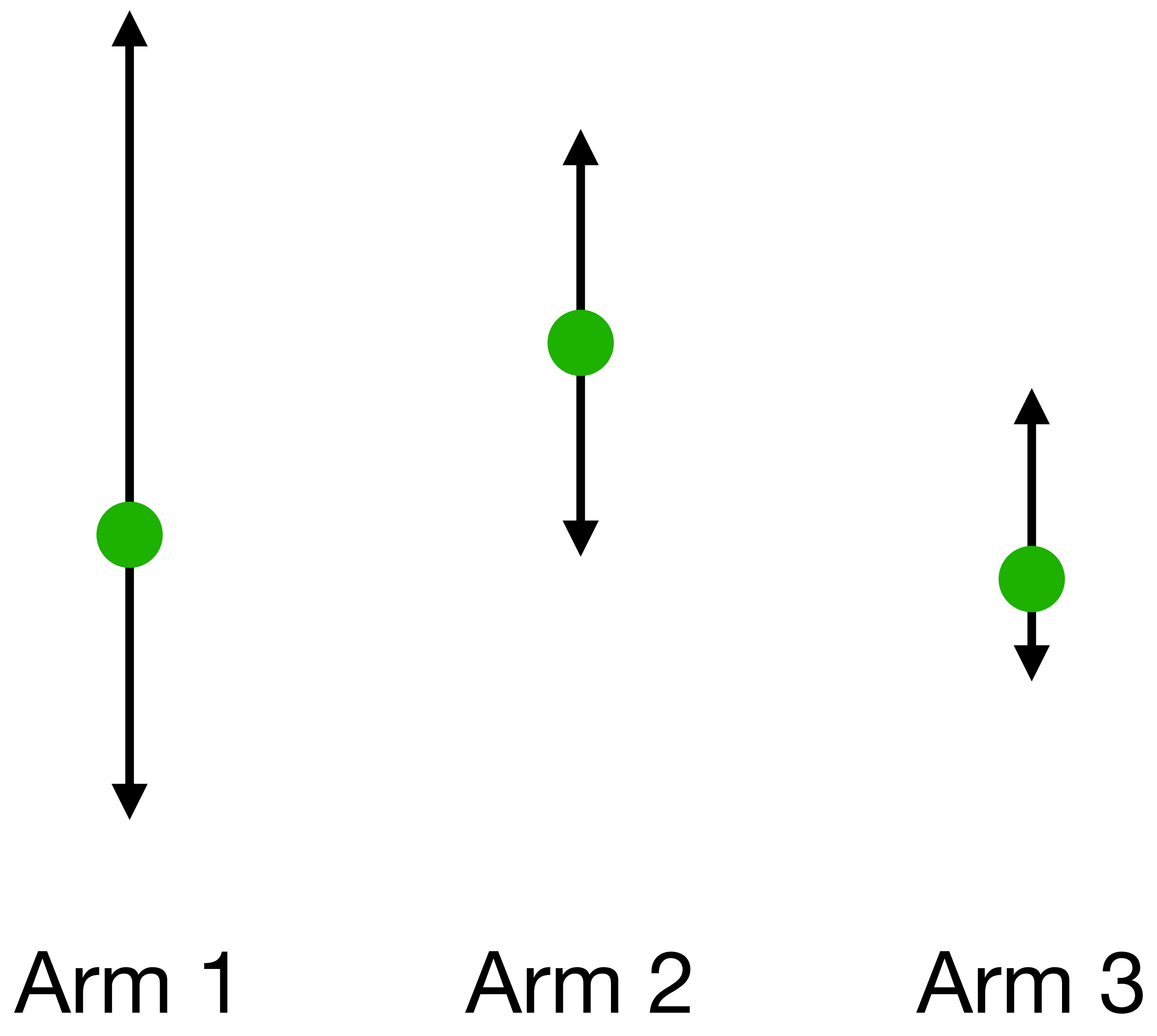
Exploration in Tabular MDPs

Kaiwen Wang and Wen Sun

CS 6789: Foundations of Reinforcement Learning

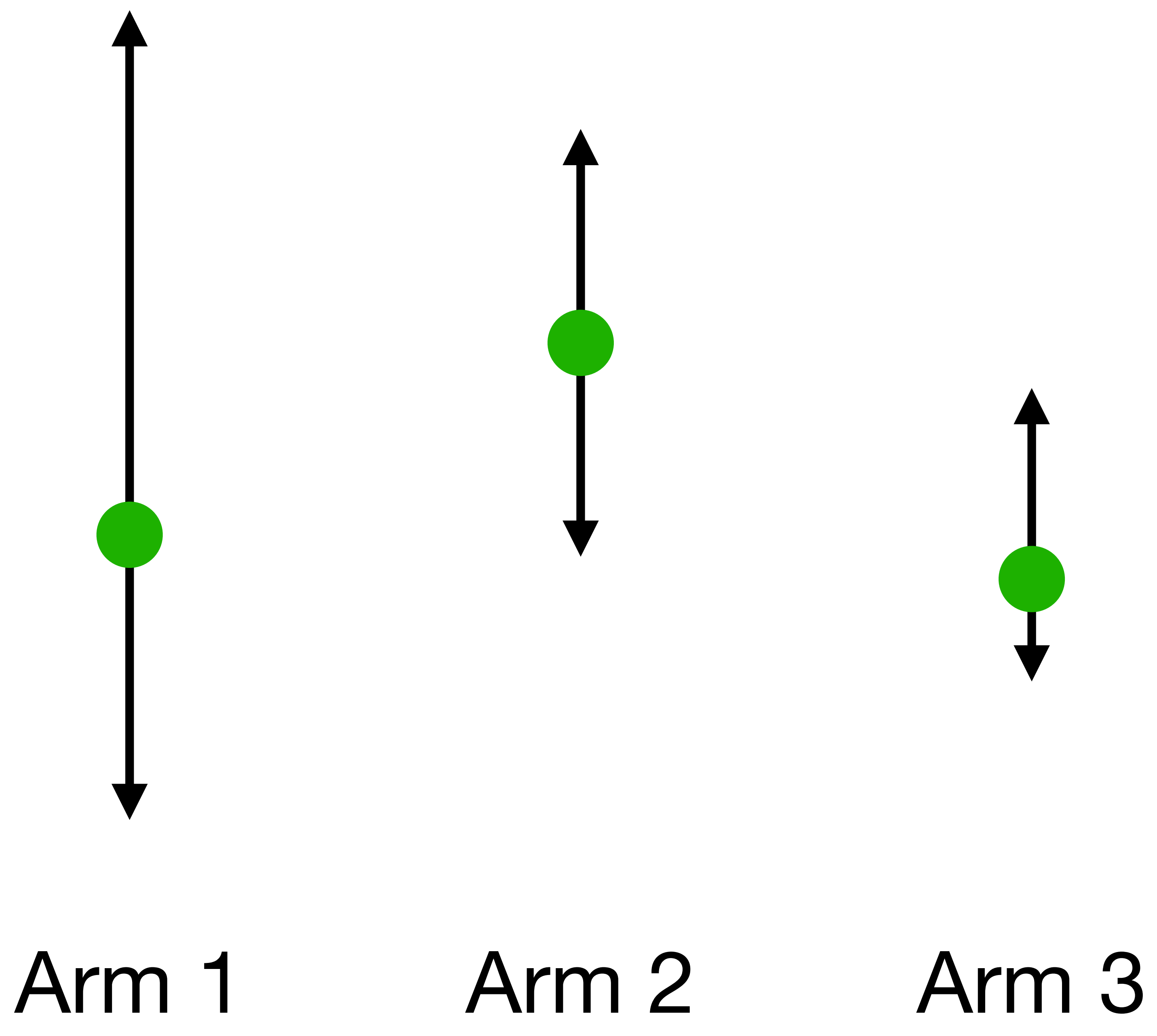
Recap:

Multi-armed Bandits and UCB Algorithm



Recap:

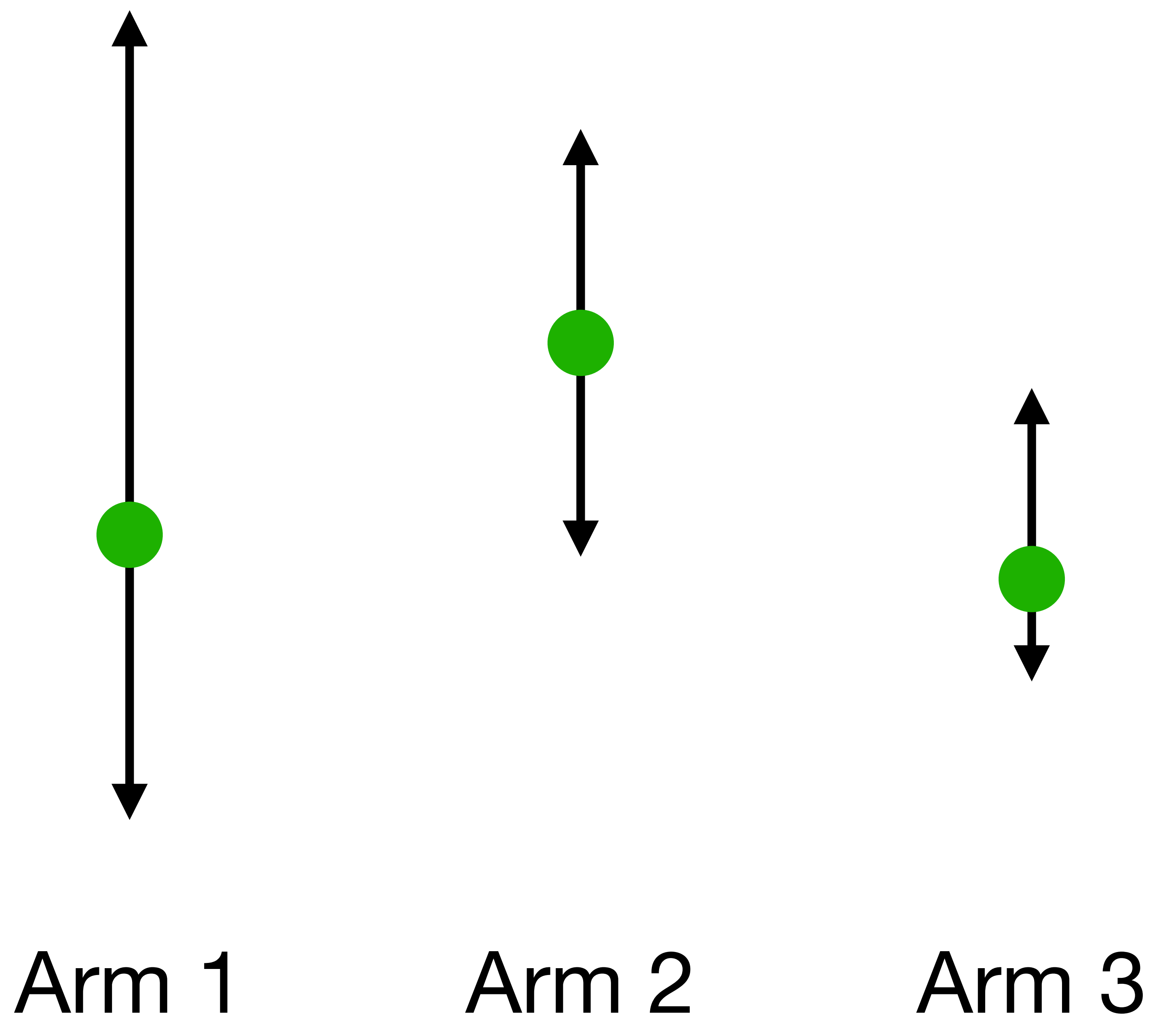
Multi-armed Bandits and UCB Algorithm



$$a^n := \arg\max_a \{\hat{\mu}^n(a) + \sqrt{\ln(KN/\delta)/N^n(a)}\}$$

Recap:

Multi-armed Bandits and UCB Algorithm

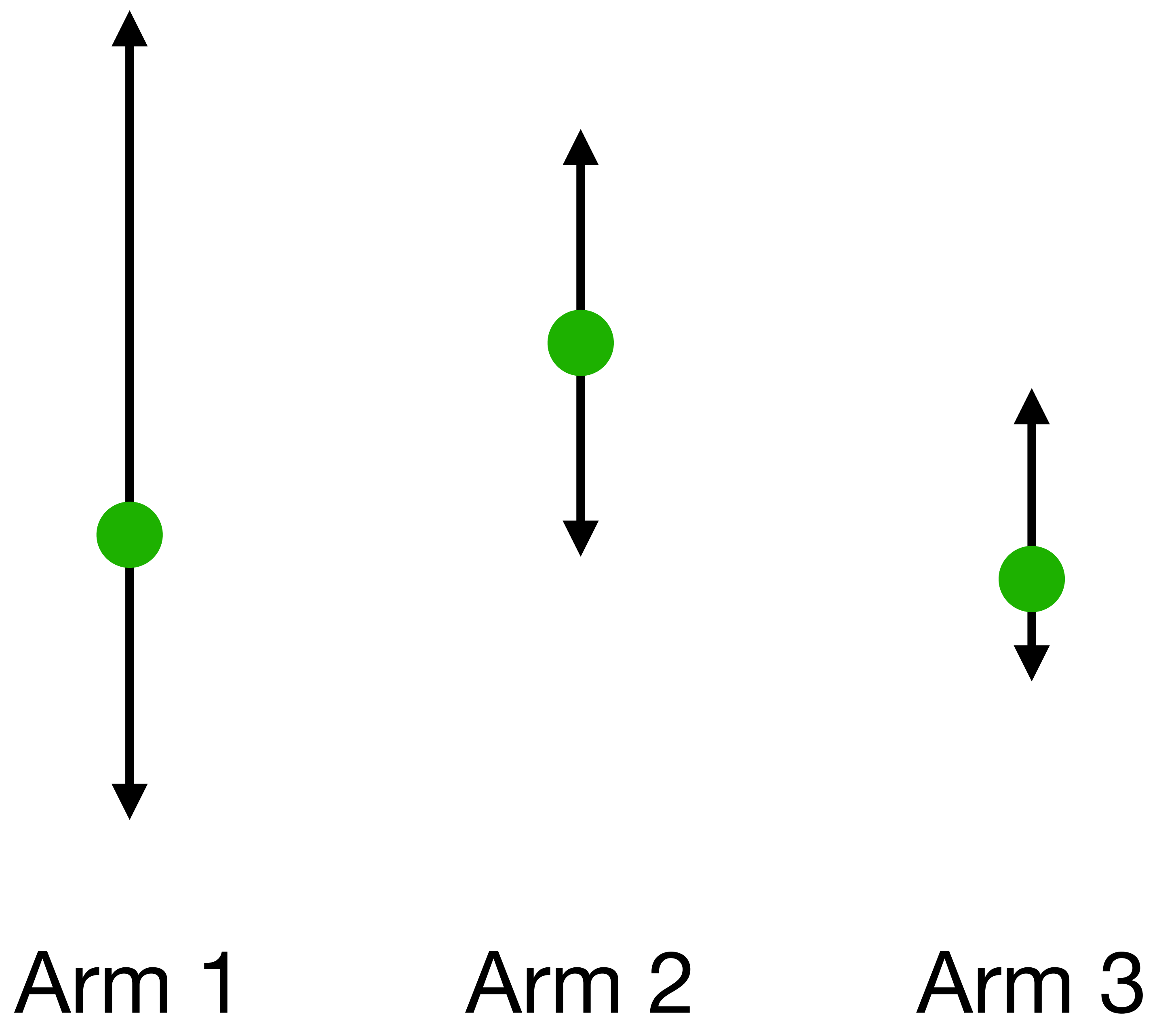


$$a^n := \arg\max_a \{\hat{\mu}^n(a) + \sqrt{\ln(KN/\delta)/N^n(a)}\}$$

$$\mathbb{E}\left[N\mu(a^\star) - \sum_{n=1}^{N}\mu(a^n)\right] \leq \widetilde{O}(\sqrt{KN})$$

Recap:

Multi-armed Bandits and UCB Algorithm



$$a^n := \arg\max_a \{\hat{\mu}^n(a) + \sqrt{\ln(KN/\delta)/N^n(a)}\}$$

$$\mathbb{E}\left[N\mu(a^\star) - \sum_{n=1}^{N} \mu(a^n)\right] \leq \widetilde{O}(\sqrt{KN})$$

Key step in the proof:

$$\mu(a^\star) - \mu(a^n) \leq \hat{\mu}(a^n) + \sqrt{\frac{\ln(KN/\delta)}{N^n(a_n)}} - \mu(a^n)$$

"Optimism"

$$\mu(a^\star) \leq UCB(a^\star) \leq UCB(a^n)$$

Today: Efficient Learning in Finite Horizon tabular MDPs

Finite horizon episode (time-dependent) discrete MDP $\mathscr{M} = \left\{ \{r_h\}_{h=0}^{H-1}, P, H, \mu, S, A \right\}$

Today: Efficient Learning in Finite Horizon tabular MDPs

Finite horizon episode (time-dependent) discrete MDP $\mathscr{M} = \left\{ \{r_h\}_{h=0}^{H-1}, P, H, \mu, S, A \right\}$

Only reset from μ: we assume it's a delta distribution, all mass at a fixed s_0

Unknown Transition P (for simplicity assume reward is known)

Today: Efficient Learning in Finite Horizon tabular MDPs

Finite horizon episode (time-dependent) discrete MDP $\mathcal{M} = \left\{ \{r_h\}_{h=0}^{H-1}, P, H, \mu, S, A \right\}$

Only reset from μ: we assume it's a delta distribution, all mass at a fixed s_0

Unknown Transition P (for simplicity assume reward is known)

Different from the Generative Model Setting!

Today: Efficient Learning in Finite Horizon tabular MDPs

Finite horizon episode (time-dependent) discrete MDP $\mathcal{M} = \left\{ \{r_h\}_{h=0}^{H-1}, P, H, \mu, S, A \right\}$

unknown

known.

Only reset from μ: we assume it's a delta distribution, all mass at a fixed s_0

Unknown Transition P (for simplicity assume reward is known)

Different from the Generative Model Setting!

EXPLORATION!

Why we need strategic exploration?

Initialization: s_0

Length of chain is H

Why we need strategic exploration?

Initialization: s_0

Length of chain is H

Probability of random walk hitting reward 1 is $(1/3)^H$

Learning Protocol

Learning Protocol

1. Learner initializes a policy π^1

Learning Protocol

1. Learner initializes a policy π^1

2. At episode n, learner executes π^n:
$\{s_h^n, a_h^n, r_h^n\}_{h=0}^{H-1}$, with $a_h^n = \pi^n(s_h^n), r_h^n = r(s_h^n, a_h^n),\ s_{h+1}^n \sim P(\cdot \mid s_h^n, a_h^n)$

Learning Protocol

1. Learner initializes a policy π^1

2. At episode n, learner executes π^n:

$$\{s_h^n, a_h^n, r_h^n\}_{h=0}^{H-1}, \text{with } a_h^n = \pi^n(s_h^n), r_h^n = r(s_h^n, a_h^n),\ s_{h+1}^n \sim P(\cdot \mid s_h^n, a_h^n)$$

3. Learner updates policy to π^{n+1} using all prior information

Learning Protocol

1. Learner initializes a policy π^1

2. At episode n, learner executes π^n:

$\{s_h^n, a_h^n, r_h^n\}_{h=0}^{H-1}$, with $a_h^n = \pi^n(s_h^n), r_h^n = r(s_h^n, a_h^n), s_{h+1}^n \sim P(\cdot \mid s_h^n, a_h^n)$

$\mu = \delta(s_0)$

3. Learner updates policy to π^{n+1} using all prior information

Performance measure: REGRET

$V^\star = \max_\pi V^\pi$

all possible

$$\text{Regret}(N) := \mathbb{E}\left[\sum_{n=1}^{N}\left(V^\star - V^{\pi^n}\right)\right] = \text{poly}(S, A, H)\sqrt{N}$$

Notations for Today

$$\mathbb{E}_{s' \sim P(\cdot|s,a)}\left[f(s')\right] := P(\cdot \,|\, s, a) \cdot f$$

$d_h^\pi(s,a)$: state-action distribution induced by π at time step h (i.e., probability of π visiting (s,a) at time step h starting from s_0)

$$\pi = \{\pi_0, \ldots, \pi_{H-1}\}$$

Outline for Today

1a. Attempt 1: Treat MDP as a Multi-armed bandit problem and run UCB

1b. Attempt 2: The Upper Confidence Bound Value Iteration Algorithm (UCB-VI)

2. UCB-VI's regret bound and the analysis

Attempt 1: Convert it to MAB and Run UCB

Q: given a discrete MDP, how many unique policies we have?

Attempt 1: Convert it to MAB and Run UCB

Q: given a discrete MDP, how many unique policies we have?

$$\left(A^S\right)^H$$

$\pi_h : S \to A$

$\pi = [\pi_0 \dots \pi_{H-1}]$

Attempt 1: Convert it to MAB and Run UCB

Q: given a discrete MDP, how many unique policies we have?

$$\left(A^S\right)^H$$

So treating each policy as an "arm", and runn UCB gives us $O(\sqrt{A^{SH} K})$

$\tilde{O}(\sqrt{A \cdot N})$

Attempt 1: Convert it to MAB and Run UCB

Q: given a discrete MDP, how many unique policies we have?

$$\left(A^S\right)^H$$

So treating each policy as an "arm", and runn UCB gives us $O(\sqrt{A^{SH}K})$

Key lesson: shouldn't treat policies as independent arms — they do share information

Outline for Today

✓ 1. Attempt 1: Treat MDP as a Multi-armed bandit problem and run UCB

1. Attempt 2: The Upper Confidence Bound Value Iteration Algorithm (UCB-VI)

2. UCB-VI's regret bound and the analysis

P

UCBVI: Optimistic Model-based Learning

Inside iteration n :

UCBVI: Optimistic Model-based Learning

Inside iteration n :

Use all previous data to estimate transitions $\widehat{P}^n \approx P$

UCBVI: Optimistic Model-based Learning

Inside iteration n :

Use all previous data to estimate transitions $\widehat{P}^n$

Design reward bonus $b_h^n(s, a), \forall s, a, h$

UCBVI: Optimistic Model-based Learning

Inside iteration n :

Use all previous data to estimate transitions $\widehat{P}^n$

Design reward bonus $b_h^n(s,a), \forall s, a, h$

Optimistic planning with learned model: $\pi^n = \text{Value-Iter}\left(\widehat{P}^n, \{r_h + b_h^n\}_{h=1}^{H-1}\right)$ dynamic programming

UCBVI: Optimistic Model-based Learning

Inside iteration n :

Use all previous data to estimate transitions $\widehat{P}^n$

Design reward bonus $b_h^n(s, a), \forall s, a, h$

Optimistic planning with learned model: $\pi^n = \text{Value-Iter}\left(\widehat{P}^n, \{r_h + b_h^n\}_{h=1}^{H-1}\right)$

Collect a new trajectory by executing π^n in the real world P starting from s_0

UCBVI—Part 1: Model Estimation

Let us consider the **very beginning** of episode n:

$$\mathscr{D}_h^n = \{s_h^i, a_h^i, s_{h+1}^i\}_{i=1}^{n-1}, \forall h$$

UCBVI—Part 1: Model Estimation

Let us consider the **very beginning** of episode n:

$$\mathcal{D}_h^n = \{s_h^i, a_h^i, s_{h+1}^i\}_{i=1}^{n-1}, \forall h$$

Let's also maintain some statistics using these datasets:

UCBVI—Part 1: Model Estimation

Let us consider the **very beginning** of episode n:

$$\mathcal{D}_h^n = \{s_h^i, a_h^i, s_{h+1}^i\}_{i=1}^{n-1}, \forall h$$

Let's also maintain some statistics using these datasets:

$$N^n(s,a) = \sum_{i=1}^{n-1} \sum_h \mathbf{1}\{(s_h^i, a_h^i) = (s,a)\}, \qquad N^n(s,a,s') = \sum_{i=1}^{n-1} \sum_h \mathbf{1}\{(s_h^i, a_h^i, s_{h+1}^i) = (s,a,s')\}.$$

↑ # visits to (s,a)

↑ # visits (s,a,s')

UCBVI—Part 1: Model Estimation

Let us consider the **very beginning** of episode n:

$$\mathscr{D}_h^n = \{s_h^i, a_h^i, s_{h+1}^i\}_{i=1}^{n-1}, \forall h$$

Let's also maintain some statistics using these datasets:

$$N^n(s,a) = \sum_{i=1}^{n-1} \sum_h \mathbf{1}\{(s_h^i, a_h^i) = (s,a)\}, \qquad N^n(s,a,s') = \sum_{i=1}^{n-1} \sum_h \mathbf{1}\{(s_h^i, a_h^i, s_{h+1}^i) = (s,a,s')\}.$$

Estimate model $\widehat{P}^n(s' \mid s,a), \forall s,a,s'$:

$$\widehat{P}^n(s' \mid s,a) = \frac{N^n(s,a,s')}{N^n(s,a)}$$

MLE

UCBVI—Part 2: Reward Bonus Design and Value Iteration

Let us consider the very beginning of episode n:

$$\mathcal{D}_h^n = \{s_h^i, a_h^i, s_{h+1}^i\}_{i=1}^{n-1}, \quad N^n(s,a) = \sum_{i=1}^{n-1} \sum_h \mathbf{1}\{(s_h^i, a_h^i) = (s,a)\}, \forall s, a,$$

UCBVI—Part 2: Reward Bonus Design and Value Iteration

Let us consider the very beginning of episode n:

$$\mathcal{D}_h^n = \{s_h^i, a_h^i, s_{h+1}^i\}_{i=1}^{n-1}, \quad N^n(s,a) = \sum_{i=1}^{n-1} \sum_h \mathbf{1}\{(s_h^i, a_h^i) = (s,a)\}, \forall s, a,$$

$$b_h^n(s,a) = cH\sqrt{\frac{\ln(SAHN/\delta)}{N^n(s,a)}}$$

UCBVI—Part 2: Reward Bonus Design and Value Iteration

Let us consider the very beginning of episode n:

$$\mathcal{D}_h^n = \{s_h^i, a_h^i, s_{h+1}^i\}_{i=1}^{n-1}, \quad N^n(s,a) = \sum_{i=1}^{n-1} \sum_h \mathbf{1}\{(s_h^i, a_h^i) = (s,a)\}, \forall s, a,$$

$$b_h^n(s,a) = cH\sqrt{\frac{\ln(SAHN/\delta)}{N^n(s,a)}}$$

Encourage to explore new state-actions

UCBVI—Part 2: Reward Bonus Design and Value Iteration

Let us consider the very beginning of episode n:

$$\mathcal{D}_h^n = \{s_h^i, a_h^i, s_{h+1}^i\}_{i=1}^{n-1}, \quad N^n(s,a) = \sum_{i=1}^{n-1} \sum_h \mathbf{1}\{(s_h^i, a_h^i) = (s,a)\}, \forall s, a,$$

$$b_h^n(s,a) = cH\sqrt{\frac{\ln(SAHN/\delta)}{N^n(s,a)}}$$

Encourage to explore new state-actions

Value Iteration (aka DP) at episode n using $\widehat{P}^n$ **and** $\{r_h + b_h^n\}_h$

UCBVI—Part 2: Reward Bonus Design and Value Iteration

Let us consider the very beginning of episode n:

$$\mathscr{D}_h^n = \{s_h^i, a_h^i, s_{h+1}^i\}_{i=1}^{n-1}, \quad N^n(s,a) = \sum_{i=1}^{n-1} \sum_h \mathbf{1}\{(s_h^i, a_h^i) = (s,a)\}, \forall s, a,$$

$$b_h^n(s,a) = cH\sqrt{\frac{\ln\left(SAHN/\delta\right)}{N^n(s,a)}}$$

Encourage to explore new state-actions

Value Iteration (aka DP) at episode n using $\widehat{P}^n$ **and** $\{r_h + b_h^n\}_h$

$$\widehat{V}_H^n(s) = 0, \forall s$$

UCBVI—Part 2: Reward Bonus Design and Value Iteration

Let us consider the very beginning of episode n:

$$\mathscr{D}_h^n = \{s_h^i, a_h^i, s_{h+1}^i\}_{i=1}^{n-1}, \quad N^n(s,a) = \sum_{i=1}^{n-1} \sum_h \mathbf{1}\{(s_h^i, a_h^i) = (s,a)\}, \forall s, a,$$

$$b_h^n(s,a) = cH\sqrt{\frac{\ln(SAHN/\delta)}{N^n(s,a)}}$$

Encourage to explore new state-actions

Value Iteration (aka DP) at episode n using $\widehat{P}^n$ **and** $\{r_h + b_h^n\}_h$

$$\widehat{V}_H^n(s) = 0, \forall s \qquad \widehat{Q}_h^n(s,a) = \min\left\{ r_h(s,a) + b_h^n(s,a) + \widehat{P}^n(\cdot \mid s,a) \cdot \widehat{V}_{h+1}^n, \quad H \right\}, \forall s, a$$

UCBVI—Part 2: Reward Bonus Design and Value Iteration

Let us consider the very beginning of episode n:

$$\mathcal{D}_h^n = \{s_h^i, a_h^i, s_{h+1}^i\}_{i=1}^{n-1}, \qquad N^n(s,a) = \sum_{i=1}^{n-1} \sum_{h} \mathbf{1}\{(s_h^i, a_h^i) = (s,a)\}, \forall s, a,$$

$$b_h^n(s,a) = cH\sqrt{\frac{\ln\left(SAHN/\delta\right)}{N^n(s,a)}}$$

Encourage to explore new state-actions

Value Iteration (aka DP) at episode n using $\widehat{P}^n$ **and** $\{r_h + b_h^n\}_h$

$$\widehat{V}_H^n(s) = 0, \forall s \qquad \widehat{Q}_h^n(s,a) = \min\left\{ r_h(s,a) + b_h^n(s,a) + \widehat{P}^n(\cdot \mid s,a) \cdot \widehat{V}_{h+1}^n, \quad H \right\}, \forall s, a$$

$$\widehat{V}_h^n(s) = \max_a \widehat{Q}_h^n(s,a), \quad \pi_h^n(s) = \arg\max_a \widehat{Q}_h^n(s,a), \forall s$$

UCBVI—Part 2: Reward Bonus Design and Value Iteration

Let us consider the very beginning of episode n:

$$\mathscr{D}_h^n = \{s_h^i, a_h^i, s_{h+1}^i\}_{i=1}^{n-1}, \quad N^n(s,a) = \sum_{i=1}^{n-1} \sum_h \mathbf{1}\{(s_h^i, a_h^i) = (s,a)\}, \forall s, a,$$

$$b_h^n(s,a) = cH\sqrt{\frac{\ln(SAHN/\delta)}{N^n(s,a)}}$$

Encourage to explore new state-actions

$Q^\pi \in [0, H]$

$H-h$

Value Iteration (aka DP) at episode n using $\widehat{P}^n$ **and** $\{r_h + b_h^n\}_h$

$$\widehat{V}_H^n(s) = 0, \forall s \qquad \widehat{Q}_h^n(s,a) = \min\left\{ r_h(s,a) + b_h^n(s,a) + \widehat{P}^n(\cdot \mid s,a) \cdot \widehat{V}_{h+1}^n, \; H \right\}, \forall s, a$$

$$\widehat{V}_h^n(s) = \max_a \widehat{Q}_h^n(s,a), \quad \pi_h^n(s) = \arg\max_a \widehat{Q}_h^n(s,a), \forall s$$

$$\left\| \widehat{V}_h^n \right\|_\infty \leq H, \forall h, n$$

UCBVI: Put All Together

For $n = 1 \to N$:

1. Set $N^n(s,a) = \sum_{i=1}^{n-1} \sum_h \mathbf{1}\{(s_h^i, a_h^i) = (s,a)\}, \forall s, a$

2. Set $N^n(s,a,s') = \sum_{i=1}^{n-1} \sum_h \mathbf{1}\{(s_h^i, a_h^i, s_{h+1}^i) = (s,a,s')\}, \forall s, a, s'$

3. Estimate model: $\widehat{P}^n(s' \mid s,a) = \dfrac{N^n(s,a,s')}{N^n(s,a)}, \forall s, a, s'$

4. Plan: $\pi^n = VI\left(\widehat{P}^n, \{r_h + b_h^n\}_h\right)$, with $b_h^n(s,a) = cH\sqrt{\dfrac{\ln(SAHN/\delta)}{N^n(s,a)}}$

5. Execute π^n : $\{s_0^n, a_0^n, r_0^n, \ldots, s_{H-1}^n, a_{H-1}^n, r_{H-1}^n, s_H^n\}$

Outline for Today

1. Attempt 1: Treat MDP as a Multi-armed bandit problem and run UCB

1. Attempt 2: The Upper Confidence Bound Value Iteration Algorithm (UCB-VI)

2. UCB-VI's regret bound and the analysis

Theorem: UCBVI Regret Bound

With probability $1 - \delta$, we have

$$\text{Regret}_N := \sum_{n=1}^{N} \left(V^\star - V^{\pi^n}\right) \leq \widetilde{O}\left(H^{1.5}\sqrt{S^2 A N \log(1/\delta)}\right)$$

Theorem: UCBVI Regret Bound

With probability $1-\delta$, we have

$$\text{Regret}_N := \sum_{n=1}^{N} \left(V^\star - V^{\pi^n}\right) \leq \widetilde{O}\left(H^{1.5}\sqrt{S^2 A N \log(1/\delta)}\right)$$

Remarks:

High probability regret implies bound on the expected regret by integrating over δ.

Theorem: UCBVI Regret Bound

With probability $1 - \delta$, we have

$$\text{Regret}_N := \sum_{n=1}^{N} \left(V^\star - V^{\pi^n}\right) \leq \widetilde{O}\left(H^{1.5}\sqrt{S^2 AN \log(1/\delta)}\right)$$

$$\mathbb{E}X = \int \Pr(X > t)\, dt$$

if $X \geq 0$

Remarks:

High probability regret implies bound on the expected regret by integrating over δ.

Dependency on H and S are suboptimal; but the **same** algorithm can achieve $H^{1.5}\sqrt{SAN}$ in the leading term [Azar et.al 17 ICML, and the book Chapter 7]

Outline of Proof

Bonus $b_h^n(s,a)$ is related to $\left(\left(\widehat{P}^n(\cdot \mid s,a) - P(\cdot \mid s,a)\right) \cdot V_{h+1}^\star\right)$

Outline of Proof

Bonus $b_h^n(s,a)$ is related to $\left(\left(\widehat{P}^n(\cdot \mid s,a) - P(\cdot \mid s,a) \right) \cdot V^\star_{h+1} \right)$

VI with bonus inside the learned model gives optimism, i.e., $\widehat{V}_h^n(s) \geq V_h^\star(s), \forall h, n, s, a$

Outline of Proof

Bonus $b_h^n(s,a)$ is related to $\left(\left(\widehat{P}^n(\cdot \mid s,a) - P(\cdot \mid s,a)\right) \cdot V_{h+1}^{\star}\right)$

VI with bonus inside the learned model gives optimism, i.e., $\widehat{V}_h^n(s) \geq V_h^{\star}(s), \forall h, n, s, a$

Upper bound per-episode regret: $V_0^{\star}(s_0) - V_0^{\pi^n}(s_0) \leq \widehat{V}_0^n(s_0) - V_0^{\pi^n}(s_0)$

Outline of Proof

Bonus $b_h^n(s, a)$ is related to $\left(\left(\widehat{P}^n(\cdot \mid s, a) - P(\cdot \mid s, a) \right) \cdot V_{h+1}^\star \right)$

VI with bonus inside the learned model gives optimism, i.e., $\widehat{V}_h^n(s) \geq V_h^\star(s), \forall h, n, s, a$

Upper bound per-episode regret: $V_0^\star(s_0) - V_0^{\pi^n}(s_0) \leq \widehat{V}_0^n(s_0) - V_0^{\pi^n}(s_0)$

Apply simulation lemma: $\widehat{V}_0^n(s_0) - V^{\pi^n}(s_0)$

1. Model Error using Hoeffing's inequality & Union Bound

$$\widehat{P}^n(s'|s,a) = \frac{N^n(s,a,s')}{N^n(s,a)}, \forall s, a, s'$$

1. Model Error using Hoeffing's inequality & Union Bound

$$\widehat{P}^n(s'|s,a) = \frac{N^n(s,a,s')}{N^n(s,a)}, \forall s, a, s'$$

Given a fixed function $f : S \mapsto [0,H]$, w/ prob $1-\delta$:

$$\left| \left(\widehat{P}^n(\cdot|s,a) - P(\cdot|s,a) \right)^\top f \right| \leq O(H\sqrt{\ln(SAHN/\delta)/N^n(s,a)}), \forall s, a, N$$

1. Model Error using Hoeffing's inequality & Union Bound

$$\widehat{P}^n(s'|s,a) = \frac{N^n(s,a,s')}{N^n(s,a)}, \forall s, a, s'$$

Given a fixed function $f : S \mapsto [0,H]$, w/ prob $1-\delta$:

$$\left| \left(\widehat{P}^n(\cdot|s,a) - P(\cdot|s,a) \right)^\top f \right| \leq O(H\sqrt{\ln(SAHN/\delta)/N^n(s,a)}), \forall s, a, N$$

Bonus $b_h^n(s,a)$

1. Model Error using Hoeffing's inequality & Union Bound

$$\widehat{P}^n(s'|s,a) = \frac{N^n(s,a,s')}{N^n(s,a)}, \forall s, a, s'$$

Given a fixed function $f : S \mapsto [0,H]$, w/ prob $1 - \delta$:

$$\left| \left(\widehat{P}^n(\cdot|s,a) - P(\cdot|s,a) \right)^\top f \right| \leq O(H\sqrt{\ln(SAHN/\delta)/N^n(s,a)}), \forall s, a, N$$

Bonus $b_h^n(s,a)$

From now on, assume this event being true

1. Model Error using Hoeffing's inequality & Union Bound

$$\widehat{P}^n(s'|s,a) = \frac{N^n(s,a,s')}{N^n(s,a)}, \forall s,a,s'$$

Given a fixed function $f: S \mapsto [0,H]$, w/ prob $1-\delta$:

$$\left| \left(\widehat{P}^n(\cdot|s,a) - P(\cdot|s,a) \right)^\top f \right| \leq O(H\sqrt{\ln(SAHN/\delta)/N^n(s,a)}), \forall s,a,N$$

Bonus $b_h^n(s,a)$

From now on, assume this event being true

Intuition:

1. Model Error using Hoeffing's inequality & Union Bound

$$\widehat{P}^n(s'|s,a) = \frac{N^n(s,a,s')}{N^n(s,a)}, \forall s,a,s'$$

Given a fixed function $f : S \mapsto [0,H]$, w/ prob $1-\delta$:

$$\left|\left(\widehat{P}^n(\cdot|s,a) - P(\cdot|s,a)\right)^{\top} f\right| \leq O(H\sqrt{\ln(SAHN/\delta)/N^n(s,a)}), \forall s,a,N$$

Bonus $b_h^n(s,a)$

$\|f\|_\infty \leq H$

$\hat{P}\cdot f = \mathbb{E}_{\hat{P}}[f(s')]$

$P\cdot f = \mathbb{E}_{P}[f(s')]$

From now on, assume this event being true

Intuition:

1. Assume for some i, $s_h^i = s, a_h^i = a$, then $f(s_{h+1}^i)$ is an unbiased estimate of $\mathbb{E}_{s'\sim P_h(\cdot|s,a)} f(s')$

1. Model Error using Hoeffing's inequality & Union Bound

$$\widehat{P}^n(s'|s,a) = \frac{N^n(s,a,s')}{N^n(s,a)}, \forall s,a,s'$$

Given a fixed function $f : S \mapsto [0,H]$, w/ prob $1-\delta$:

$$\left| \left(\widehat{P}^n(\cdot|s,a) - P(\cdot|s,a) \right)^\top f \right| \leq O(H\sqrt{\ln(SAHN/\delta)/N^n(s,a)}), \forall s,a,N$$

Bonus $b_h^n(s,a)$

From now on, assume this event being true

Intuition:

1. Assume for some i, $s_h^i = s, a_h^i = a$, then $f(s_{h+1}^i)$ is an unbiased estimate of $\mathbb{E}_{s' \sim P_h(\cdot|s,a)} f(s')$

2. Note $\widehat{P}^n(\cdot|s,a) \cdot f = \frac{1}{N^n(s,a)} \sum_{i=1}^{n-1} \sum_h \mathbf{1}[(s_h^i, a_h^i) = (s,a)] f(s_{h+1}^i)$

2. Proving Optimism via Induction

Lemma [Optimism]: $\widehat{V}_h^n(s) \geq V_h^\star(s), \forall n, h, s$

Recall Bonus-enhanced Value Iteration at episode n:

$$\widehat{V}_H^n(s) = 0, \quad \widehat{Q}_h^n(s,a) = \min\left\{ r_h(s,a) + b_h^n(s,a) + \widehat{P}^n(\cdot \mid s,a) \cdot \widehat{V}_{h+1}^n, H \right\}$$

$$\widehat{V}_h^n(s) = \max_a \widehat{Q}_h^n(s,a), \quad \pi_h^n(s) = \arg\max_a \widehat{Q}_h^n(s,a), \forall s$$

2. Proving Optimism via Induction

Lemma [Optimism]: $\widehat{V}_h^n(s) \geq V_h^\star(s), \forall n, h, s$

Recall Bonus-enhanced Value Iteration at episode n:

$$\widehat{V}_H^n(s) = 0, \quad \widehat{Q}_h^n(s,a) = \min\left\{ r_h(s,a) + b_h^n(s,a) + \widehat{P}^n(\cdot \mid s,a) \cdot \widehat{V}_{h+1}^n, H \right\}$$

$$\widehat{V}_h^n(s) = \max_a \widehat{Q}_h^n(s,a), \quad \pi_h^n(s) = \arg\max_a \widehat{Q}_h^n(s,a), \forall s$$

Inductive hypothesis: $\widehat{V}_{h+1}^n(s) \geq V_{h+1}^\star(s), \quad \forall s$

2. Proving Optimism via Induction

Lemma [Optimism]: $\widehat{V}_h^n(s) \geq V_h^\star(s), \forall n, h, s$

Recall Bonus-enhanced Value Iteration at episode n:

$$\widehat{V}_H^n(s) = 0, \quad \widehat{Q}_h^n(s,a) = \min\left\{ r_h(s,a) + b_h^n(s,a) + \widehat{P}^n(\cdot \mid s,a) \cdot \widehat{V}_{h+1}^n, H \right\}$$

$$\widehat{V}_h^n(s) = \max_a \widehat{Q}_h^n(s,a), \quad \pi_h^n(s) = \arg\max_a \widehat{Q}_h^n(s,a), \forall s$$

Inductive hypothesis: $\widehat{V}_{h+1}^n(s) \geq V_{h+1}^\star(s), \quad \forall s$

$$\widehat{Q}_h^n(s,a) - Q_h^\star(s,a) = r_h(s,a) + b_h^n(s,a) + \widehat{P}^n(\cdot \mid s,a) \cdot \widehat{V}_{h+1}^n - r_h(s,a) - P(\cdot \mid s,a) \cdot V_{h+1}^\star$$

2. Proving Optimism via Induction

Lemma [Optimism]: $\widehat{V}_h^n(s) \geq V_h^\star(s), \forall n, h, s$

Recall Bonus-enhanced Value Iteration at episode n:

$$\widehat{V}_H^n(s) = 0, \quad \widehat{Q}_h^n(s,a) = \min\left\{ r_h(s,a) + b_h^n(s,a) + \widehat{P}^n(\cdot \mid s,a) \cdot \widehat{V}_{h+1}^n, H \right\}$$

$$\widehat{V}_h^n(s) = \max_a \widehat{Q}_h^n(s,a), \quad \pi_h^n(s) = \arg\max_a \widehat{Q}_h^n(s,a), \forall s$$

Inductive hypothesis: $\widehat{V}_{h+1}^n(s) \geq V_{h+1}^\star(s), \quad \forall s$

$$\widehat{Q}_h^n(s,a) - Q_h^\star(s,a) = r_h(s,a) + b_h^n(s,a) + \widehat{P}^n(\cdot \mid s,a) \cdot \widehat{V}_{h+1}^n - r_h(s,a) - P(\cdot \mid s,a) \cdot V_{h+1}^\star$$

$\geq V^\star_{h+1}(s)$

$$\geq b_h^n(s,a) + \widehat{P}^n(\cdot \mid s,a) \cdot V_{h+1}^\star - P(\cdot \mid s,a) \cdot V_{h+1}^\star$$

2. Proving Optimism via Induction

Lemma [Optimism]: $\widehat{V}_h^n(s) \geq V_h^\star(s), \forall n, h, s$

Recall Bonus-enhanced Value Iteration at episode n:

$$\widehat{V}_H^n(s) = 0, \quad \widehat{Q}_h^n(s,a) = \min\left\{ r_h(s,a) + b_h^n(s,a) + \widehat{P}^n(\cdot \mid s,a) \cdot \widehat{V}_{h+1}^n, H \right\}$$

$$\widehat{V}_h^n(s) = \max_a \widehat{Q}_h^n(s,a), \quad \pi_h^n(s) = \arg\max_a \widehat{Q}_h^n(s,a), \forall s$$

Inductive hypothesis: $\widehat{V}_{h+1}^n(s) \geq V_{h+1}^\star(s), \quad \forall s$

$$\widehat{Q}_h^n(s,a) - Q_h^\star(s,a) = r_h(s,a) + b_h^n(s,a) + \widehat{P}^n(\cdot \mid s,a) \cdot \widehat{V}_{h+1}^n - r_h(s,a) - P(\cdot \mid s,a) \cdot V_{h+1}^\star$$

$$\geq b_h^n(s,a) + \widehat{P}^n(\cdot \mid s,a) \cdot V_{h+1}^\star - P(\cdot \mid s,a) \cdot V_{h+1}^\star$$

$$= b_h^n(s,a) + \left(\widehat{P}^n(\cdot \mid s,a) - P(\cdot \mid s,a) \right) \cdot V_{h+1}^\star$$

$\in [0, H]$

$\leq H\sqrt{\frac{1}{N}} = b_h^n$

2. Proving Optimism via Induction

Lemma [Optimism]: $\widehat{V}^n_h(s) \geq V^\star_h(s), \forall n, h, s$

Recall Bonus-enhanced Value Iteration at episode n:

$$\widehat{V}^n_H(s) = 0, \quad \widehat{Q}^n_h(s,a) = \min\left\{ r_h(s,a) + b^n_h(s,a) + \widehat{P}^n(\cdot \mid s,a) \cdot \widehat{V}^n_{h+1}, H \right\}$$

$$\widehat{V}^n_h(s) = \max_a \widehat{Q}^n_h(s,a), \quad \pi^n_h(s) = \arg\max_a \widehat{Q}^n_h(s,a), \forall s$$

Inductive hypothesis: $\widehat{V}^n_{h+1}(s) \geq V^\star_{h+1}(s), \quad \forall s$

$$\widehat{Q}^n_h(s,a) - Q^\star_h(s,a) = r_h(s,a) + b^n_h(s,a) + \widehat{P}^n(\cdot \mid s,a) \cdot \widehat{V}^n_{h+1} - r_h(s,a) - P(\cdot \mid s,a) \cdot V^\star_{h+1}$$

$$\geq b^n_h(s,a) + \widehat{P}^n(\cdot \mid s,a) \cdot V^\star_{h+1} - P(\cdot \mid s,a) \cdot V^\star_{h+1}$$

$$= b^n_h(s,a) + \left(\widehat{P}^n(\cdot \mid s,a) - P(\cdot \mid s,a) \right) \cdot V^\star_{h+1}$$

$$\geq b^n_h(s,a) - b^n_h(s,a) = 0, \quad \forall s, a$$

$V^\star_h(s) - \widehat{V}^n_h(s) = \max_a Q^\star_h(s,a) - \max_a \widehat{Q}^n_h(s,a)$

$\leq \max_a (Q^\star_h - \widehat{Q}^n_h)(s,a)$

≤ 0 .

3. Upper Bounding Regret using Optimism

per-episode regret $:= V_0^\star(s_0) - V_0^{\pi_n}(s_0) \leq \widehat{V}_0^n(s_0) - V_0^{\pi_n}(s_0)$

This is something
we can control!
And this is related
to our policy π^n

4. Upper bounding Regret via Simulation Lemma

$$\widehat{V}^n_H(s) = 0, \quad \widehat{Q}^n_h(s,a) = \min\left\{ r_h(s,a) + b^n_h(s,a) + \widehat{P}^n_h(\cdot \mid s,a) \cdot \widehat{V}^n_{h+1}, H \right\}$$

$$\widehat{V}^n_h(s) = \max_a \widehat{Q}^n_h(s,a), \quad \pi^n_h(s) = \arg\max_a \widehat{Q}^n_h(s,a), \forall s$$

Lemma [Simulation lemma]:

$$\widehat{V}^n_0(s_0) - V^{\pi^n}_0(s_0) \leq \sum_{h=0}^{H-1} \mathbb{E}_{s,a \sim d^{\pi^n}_h}\left[b^n_h(s,a) + \left(\widehat{P}^n(\cdot \mid s,a) - P(\cdot \mid s,a)\right) \cdot \widehat{V}^n_{h+1} \right]$$

Value under $\widehat{P}$, $r+b$

Value under P, r

$r+b-r=b$

difference in rewards

difference in transition

4. Upper bounding Regret via Simulation Lemma

$$\widehat{V}^n_H(s) = 0, \quad \widehat{Q}^n_h(s,a) = \min\left\{ r_h(s,a) + b^n_h(s,a) + \widehat{P}^n_h(\cdot \mid s,a) \cdot \widehat{V}^n_{h+1}, H \right\}$$

$$\widehat{V}^n_h(s) = \max_a \widehat{Q}^n_h(s,a), \quad \pi^n_h(s) = \arg\max_a \widehat{Q}^n_h(s,a), \forall s$$

Lemma [Simulation lemma]:

$$\widehat{V}^n_0(s_0) - V^{\pi^n}_0(s_0) \le \sum_{h=0}^{H-1} \mathbb{E}_{s,a\sim d^{\pi^n}_h}\left[b^n_h(s,a) + \left(\widehat{P}^n(\cdot \mid s,a) - P(\cdot \mid s,a)\right) \cdot \widehat{V}^n_{h+1} \right]$$

$$\widehat{V}^n_0(s_0) - V^{\pi^n}_0(s_0) = \widehat{Q}^n_0(s_0, \pi^n(s_0)) - Q^{\pi^n}_0(s_0, \pi^n(s_0))$$

4. Upper bounding Regret via Simulation Lemma

$$\widehat{V}^n_H(s) = 0, \quad \widehat{Q}^n_h(s,a) = \min\left\{ r_h(s,a) + b^n_h(s,a) + \widehat{P}^n_h(\cdot \mid s,a) \cdot \widehat{V}^n_{h+1}, H \right\}$$

$$\widehat{V}^n_h(s) = \max_a \widehat{Q}^n_h(s,a), \quad \pi^n_h(s) = \arg\max_a \widehat{Q}^n_h(s,a), \forall s$$

Lemma [Simulation lemma]:

$$\widehat{V}^n_0(s_0) - V^{\pi^n}_0(s_0) \leq \sum_{h=0}^{H-1} \mathbb{E}_{s,a \sim d^{\pi^n}_h} \left[b^n_h(s,a) + (\widehat{P}^n(\cdot \mid s,a) - P(\cdot \mid s,a)) \cdot \widehat{V}^n_{h+1} \right]$$

$$\widehat{V}^n_0(s_0) - V^{\pi^n}_0(s_0) = \widehat{Q}^n_0(s_0, \pi^n(s_0)) - Q^{\pi^n}_0(s_0, \pi^n(s_0))$$

$$\leq r_0(s_0, \pi^n(s_0)) + b^n_h(s_0, \pi^n(s_0)) + \widehat{P}^n(\cdot \mid s_0, \pi^n(s_0)) \cdot \widehat{V}^n_1 - r_0(s_0, \pi^n(s_0)) - P(\cdot \mid s_0, \pi^n(s_0)) \cdot V^{\pi^n}_1$$

4. Upper bounding Regret via Simulation Lemma

$$\widehat{V}^n_H(s) = 0, \quad \widehat{Q}^n_h(s,a) = \min\left\{ r_h(s,a) + b^n_h(s,a) + \widehat{P}^n_h(\cdot \mid s,a) \cdot \widehat{V}^n_{h+1}, H \right\}$$

$$\widehat{V}^n_h(s) = \max_a \widehat{Q}^n_h(s,a), \quad \pi^n_h(s) = \arg\max_a \widehat{Q}^n_h(s,a), \forall s$$

Lemma [Simulation lemma]:

$$\widehat{V}^n_0(s_0) - V^{\pi^n}_0(s_0) \leq \sum_{h=0}^{H-1} \mathbb{E}_{s,a\sim d^{\pi^n}_h}\left[b^n_h(s,a) + (\widehat{P}^n(\cdot \mid s,a) - P(\cdot \mid s,a)) \cdot \widehat{V}^n_{h+1} \right]$$

$$\widehat{V}^n_0(s_0) - V^{\pi^n}_0(s_0) = \widehat{Q}^n_0(s_0, \pi^n(s_0)) - Q^{\pi^n}_0(s_0, \pi^n(s_0))$$

$$\leq r_0(s_0, \pi^n(s_0)) + b^n_h(s_0, \pi^n(s_0)) + \widehat{P}^n(\cdot \mid s_0, \pi^n(s_0)) \cdot \widehat{V}^n_1 - r_0(s_0, \pi^n(s_0)) - P(\cdot \mid s_0, \pi^n(s_0)) \cdot V^{\pi^n}_1$$

$$= b^n_h(s_0, \pi^n(s_0)) + \widehat{P}^n(\cdot \mid s_0, \pi^n(s_0)) \cdot \widehat{V}^n_1 - P(\cdot \mid s_0, \pi^n(s_0)) \cdot V^{\pi^n}_1$$

4. Upper bounding Regret via Simulation Lemma

$$\widehat{V}^n_H(s) = 0, \quad \widehat{Q}^n_h(s,a) = \min\left\{ r_h(s,a) + b^n_h(s,a) + \widehat{P}^n_h(\cdot \mid s,a) \cdot \widehat{V}^n_{h+1}, H \right\}$$

$$\widehat{V}^n_h(s) = \max_a \widehat{Q}^n_h(s,a), \quad \pi^n_h(s) = \arg\max_a \widehat{Q}^n_h(s,a), \forall s$$

Lemma [Simulation lemma]:

$$\widehat{V}^n_0(s_0) - V^{\pi^n}_0(s_0) \leq \sum_{h=0}^{H-1} \mathbb{E}_{s,a\sim d^{\pi^n}_h}\left[b^n_h(s,a) + \left(\widehat{P}^n(\cdot \mid s,a) - P(\cdot \mid s,a)\right) \cdot \widehat{V}^n_{h+1} \right]$$

$$\widehat{V}^n_0(s_0) - V^{\pi^n}_0(s_0) = \widehat{Q}^n_0(s_0, \pi^n(s_0)) - Q^{\pi^n}_0(s_0, \pi^n(s_0))$$

$$\leq r_0(s_0, \pi^n(s_0)) + b^n_h(s_0, \pi^n(s_0)) + \widehat{P}^n(\cdot \mid s_0, \pi^n(s_0)) \cdot \widehat{V}^n_1 - r_0(s_0, \pi^n(s_0)) - P(\cdot \mid s_0, \pi^n(s_0)) \cdot V^{\pi^n}_1$$

$$= b^n_h(s_0, \pi^n(s_0)) + \widehat{P}^n(\cdot \mid s_0, \pi^n(s_0)) \cdot \widehat{V}^n_1 - P(\cdot \mid s_0, \pi^n(s_0)) \cdot V^{\pi^n}_1$$

$$= b^n_h(s_0, \pi^n(s_0)) + \left(\widehat{P}^n(\cdot \mid s_0, \pi^n(s_0)) - P(\cdot \mid s_0, \pi^n(s_0)) \right) \cdot \widehat{V}^n_1 + P(\cdot \mid s_0, \pi^n(s_0)) \cdot \left(\widehat{V}^n_1 - V^{\pi^n}_1 \right)$$

$$P \cdot \widehat{V}^n_1$$

4. Upper bounding Regret via Simulation Lemma

$$\widehat{V}^n_H(s) = 0, \quad \widehat{Q}^n_h(s,a) = \min\left\{ r_h(s,a) + b^n_h(s,a) + \widehat{P}^n_h(\cdot\,|\,s,a)\cdot \widehat{V}^n_{h+1}, H \right\}$$

$$\widehat{V}^n_h(s) = \max_a \widehat{Q}^n_h(s,a), \quad \pi^n_h(s) = \arg\max_a \widehat{Q}^n_h(s,a), \forall s$$

Lemma [Simulation lemma]:

$$\widehat{V}^n_0(s_0) - V^{\pi^n}_0(s_0) \leq \sum_{h=0}^{H-1} \mathbb{E}_{s,a\sim d^{\pi^n}_h}\left[b^n_h(s,a) + (\widehat{P}^n(\cdot\,|\,s,a) - P(\cdot\,|\,s,a))\cdot \widehat{V}^n_{h+1} \right]$$

$$\widehat{V}^n_0(s_0) - V^{\pi^n}_0(s_0) = \widehat{Q}^n_0(s_0, \pi^n(s_0)) - Q^{\pi^n}_0(s_0, \pi^n(s_0))$$

$$\leq r_0(s_0,\pi^n(s_0)) + b^n_h(s_0,\pi^n(s_0)) + \widehat{P}^n(\cdot\,|\,s_0,\pi^n(s_0))\cdot \widehat{V}^n_1 - r_0(s_0,\pi^n(s_0)) - P(\cdot\,|\,s_0,\pi^n(s_0))\cdot V^{\pi^n}_1$$

$$= b^n_h(s_0,\pi^n(s_0)) + \widehat{P}^n(\cdot\,|\,s_0,\pi^n(s_0))\cdot \widehat{V}^n_1 - P(\cdot\,|\,s_0,\pi^n(s_0))\cdot V^{\pi^n}_1$$

$$= b^n_h(s_0,\pi^n(s_0)) + \left(\widehat{P}^n(\cdot\,|\,s_0,\pi^n(s_0)) - P(\cdot\,|\,s_0,\pi^n(s_0))\right)\cdot \widehat{V}^n_1 + P(\cdot\,|\,s_0,\pi^n(s_0))\cdot\left(\widehat{V}^n_1 - V^{\pi^n}_1\right)$$

$$= \sum_{h=0}^{H-1} \mathbb{E}_{s,a\sim d^{\pi^n}_h}\left[b^n_h(s,a) + (\widehat{P}^n(\cdot\,|\,s,a) - P(\cdot\,|\,s,a))\cdot \widehat{V}^n_{h+1} \right]$$

4. Upper bounding Regret via Simulation Lemma

per-episode regret $:= V_0^\star(s_0) - V_0^{\pi_n}(s_0) \leq \widehat{V}_0^n(s_0) - V_0^{\pi_n}(s_0)$

$$\leq \sum_{h=0}^{H-1} \mathbb{E}_{s,a\sim d_h^{\pi^n}} \left[b_h^n(s,a) + (\widehat{P}^n(\cdot \mid s,a) - P(\cdot \mid s,a)) \cdot \widehat{V}_{h+1}^n \right]$$

4. Upper bounding Regret via Simulation Lemma

per-episode regret $:= V_0^\star(s_0) - V_0^{\pi_n}(s_0) \leq \widehat{V}_0^n(s_0) - V_0^{\pi_n}(s_0)$

But $\widehat{V}_h^n$ is data-dependent

(this is different from $V_h^\star$) !!!

$$\leq \sum_{h=0}^{H-1} \mathbb{E}_{s,a\sim d_h^{\pi^n}} \left[b_h^n(s,a) + (\widehat{P}^n(\cdot \mid s,a) - P(\cdot \mid s,a)) \cdot \widehat{V}_{h+1}^n \right]$$

Let's do Holder's inequality

4. Upper bounding Regret via Simulation Lemma

per-episode regret $:= V_0^\star(s_0) - V_0^{\pi_n}(s_0) \leq \widehat{V}_0^n(s_0) - V_0^{\pi_n}(s_0)$

$$\leq \sum_{h=0}^{H-1} \mathbb{E}_{s,a\sim d_h^{\pi^n}}\left[b_h^n(s,a) + (\widehat{P}^n(\cdot\,|\,s,a) - P(\cdot\,|\,s,a))\cdot \widehat{V}_{h+1}^n\right]$$

But $\widehat{V}_h^n$ is data-dependent (this is different from $V_h^\star$) !!!

Let's do Holder's inequality

$$\left(\widehat{P}_h^n(\cdot\,|\,s,a) - P_h(\cdot\,|\,s,a)\right)\cdot \widehat{V}_{h+1}^n \leq \|P_h(\cdot\,|\,s,a) - \widehat{P}_h^n(\cdot\,|\,s,a)\|_1 \|\widehat{V}_{h+1}^n\|_\infty$$

4. Upper bounding Regret via Simulation Lemma

per-episode regret $:= V_0^\star(s_0) - V_0^{\pi_n}(s_0) \leq \widehat{V}_0^n(s_0) - V_0^{\pi_n}(s_0)$

But $\widehat{V}_h^n$ is data-dependent
(this is different from $V_h^\star$) !!!

$$\leq \sum_{h=0}^{H-1} \mathbb{E}_{s,a\sim d_h^{\pi^n}} \left[b_h^n(s,a) + (\widehat{P}^n(\cdot\,|\,s,a) - P(\cdot\,|\,s,a)) \cdot \widehat{V}_{h+1}^n \right]$$

Let's do Holder's inequality

$x \cdot y \leq \|x\|_1 \cdot \|y\|_\infty$ (Hölder's ineq.)

$$\left(\widehat{P}_h^n(\cdot\,|\,s,a) - P_h(\cdot\,|\,s,a)\right) \cdot \widehat{V}_{h+1}^n \leq \|P_h(\cdot\,|\,s,a) - \widehat{P}_h^n(\cdot\,|\,s,a)\|_1 \|\widehat{V}_{h+1}^n\|_\infty$$

$$\leq H\|P_h(\cdot\,|\,s,a) - \widehat{P}_h^n(\cdot\,|\,s,a)\|_1 \leq H\sqrt{\frac{S\ln(SAHN/\delta)}{N_h^n(s,a)}}, \forall s,a,h,n, \text{with prob} 1-\delta$$

4. Upper bounding Regret via Simulation Lemma

per-episode regret $:= V_0^\star(s_0) - V_0^{\pi_n}(s_0) \leq \widehat{V}_0^n(s_0) - V_0^{\pi_n}(s_0)$

But $\widehat{V}_h^n$ is data-dependent
(this is different from $V_h^\star$) !!!

$$\leq \sum_{h=0}^{H-1} \mathbb{E}_{s,a\sim d_h^{\pi^n}} \left[b_h^n(s,a) + (\widehat{P}^n(\cdot \mid s,a) - P(\cdot \mid s,a)) \cdot \widehat{V}_{h+1}^n \right]$$

Let's do Holder's inequality

$$\leq \sum_{h=0}^{H-1} \mathbb{E}_{s,a\sim d_h^{\pi^n}} \left[b_h^n(s,a) + H\sqrt{\frac{S \ln(SAHN/\delta)}{N^n(s,a)}} \right]$$

$$\left(\widehat{P}_h^n(\cdot \mid s,a) - P_h(\cdot \mid s,a) \right) \cdot \widehat{V}_{h+1}^n \leq \| P_h(\cdot \mid s,a) - \widehat{P}_h^n(\cdot \mid s,a) \|_1 \| \widehat{V}_{h+1}^n \|_\infty$$

$$\leq H \| P_h(\cdot \mid s,a) - \widehat{P}_h^n(\cdot \mid s,a) \|_1 \leq H\sqrt{\frac{S \ln(SAHN/\delta)}{N_h^n(s,a)}}, \forall s,a,h,n, \text{ with prob} 1-\delta$$

4. Upper bounding Regret via Simulation Lemma

per-episode regret := $V_0^\star(s_0) - V_0^{\pi_n}(s_0) \leq \widehat{V}_0^n(s_0) - V_0^{\pi_n}(s_0)$

But $\widehat{V}_h^n$ is data-dependent (this is different from $V_h^\star$) !!!

$$\leq \sum_{h=0}^{H-1} \mathbb{E}_{s,a\sim d_h^{\pi^n}} \left[b_h^n(s,a) + (\widehat{P}^n(\cdot \mid s,a) - P(\cdot \mid s,a)) \cdot \widehat{V}_{h+1}^n \right]$$

Let's do Holder's inequality

$$\leq \sum_{h=0}^{H-1} \mathbb{E}_{s,a\sim d_h^{\pi^n}} \left[b_h^n(s,a) + H\sqrt{\frac{S\ln(SAHN/\delta)}{N^n(s,a)}} \right]$$

$b = H\sqrt{\frac{\log \ldots}{N}}$

$$\leq 2\sum_{h=0}^{H-1} \mathbb{E}_{s,a\sim d_h^{\pi^n}} \left[H\sqrt{\frac{S\ln(SAHN/\delta)}{N^n(s,a)}} \right]$$

$$\left(\widehat{P}_h^n(\cdot \mid s,a) - P_h(\cdot \mid s,a)\right) \cdot \widehat{V}_{h+1}^n \leq \|P_h(\cdot \mid s,a) - \widehat{P}_h^n(\cdot \mid s,a)\|_1 \|\widehat{V}_{h+1}^n\|_\infty$$

$$\leq H\|P_h(\cdot \mid s,a) - \widehat{P}_h^n(\cdot \mid s,a)\|_1 \leq H\sqrt{\frac{S\ln(SAHN/\delta)}{N_h^n(s,a)}}, \forall s,a,h,n, \text{ with prob} 1-\delta$$

4. Upper bounding Regret via Simulation Lemma

per-episode regret $:= V_0^\star(s_0) - V_0^{\pi_n}(s_0) \leq \widehat{V}_0^n(s_0) - V_0^{\pi_n}(s_0)$

But $\widehat{V}_h^n$ is data-dependent (this is different from $V_h^\star$) !!!

$$\leq \sum_{h=0}^{H-1} \mathbb{E}_{s,a\sim d_h^{\pi^n}}\left[b_h^n(s,a) + (\widehat{P}^n(\cdot|s,a) - P(\cdot|s,a)) \cdot \widehat{V}_{h+1}^n\right]$$

Let's do Holder's inequality

$$\leq \sum_{h=0}^{H-1} \mathbb{E}_{s,a\sim d_h^{\pi^n}}\left[b_h^n(s,a) + H\sqrt{\frac{S\ln(SAHN/\delta)}{N^n(s,a)}}\right]$$

$$\leq 2\sum_{h=0}^{H-1} \mathbb{E}_{s,a\sim d_h^{\pi^n}}\left[H\sqrt{\frac{S\ln(SAHN/\delta)}{N^n(s,a)}}\right] = 2H\sqrt{S\ln(SAHN/\delta)}\sum_{h=0}^{H-1} \mathbb{E}_{s,a\sim d_h^{\pi^n}}\left[\sqrt{\frac{1}{N^n(s,a)}}\right]$$

$$\left(\widehat{P}_h^n(\cdot|s,a) - P_h(\cdot|s,a)\right)\cdot \widehat{V}_{h+1}^n \leq \|P_h(\cdot|s,a) - \widehat{P}_h^n(\cdot|s,a)\|_1 \|\widehat{V}_{h+1}^n\|_\infty$$

$$\leq H\|P_h(\cdot|s,a) - \widehat{P}_h^n(\cdot|s,a)\|_1 \leq H\sqrt{\frac{S\ln(SAHN/\delta)}{N_h^n(s,a)}}, \forall s,a,h,n, \text{with prob} 1-\delta$$

5. Final Step

Remember we had two failure events for bounding transitions errors.

5. Final Step

Remember we had two failure events for bounding transitions errors.

$$\text{Regret}_N = \sum_{n=1}^{N} \left(V_0^{\star}(s_0) - V_0^{\pi^n}(s_0) \right) \leq 2H\sqrt{S\ln(SAHN/\delta)} \sum_{n=1}^{N} \sum_{h=0}^{H-1} \mathbb{E}_{s,a\sim d_h^{\pi^n}} \left[\sqrt{\frac{1}{N^n(s,a)}} \right]$$

5. Final Step

Remember we had two failure events for bounding transitions errors.

$$\text{Regret}_N = \sum_{n=1}^{N} \left(V_0^\star(s_0) - V_0^{\pi^n}(s_0) \right) \leq 2H\sqrt{S \ln(SAHN/\delta)} \sum_{n=1}^{N} \sum_{h=0}^{H-1} \mathbb{E}_{s,a \sim d_h^{\pi^n}} \left[\sqrt{\frac{1}{N^n(s,a)}} \right]$$

$$\leq 4H\sqrt{S \ln(SAHN/\delta)} \left(\sum_{n,h} \sqrt{\frac{1}{N^n(s_h^n, a_h^n)}} + H \log(N/\delta) \right)$$

(Azuma's ineq.)

5. Final Step

Remember we had two failure events for bounding transitions errors.

$$\text{Regret}_N = \sum_{n=1}^{N} \left(V_0^\star(s_0) - V_0^{\pi^n}(s_0) \right) \leq 2H\sqrt{S \ln(SAHN/\delta)} \sum_{n=1}^{N} \sum_{h=0}^{H-1} \mathbb{E}_{s,a\sim d_h^{\pi^n}} \left[\sqrt{\frac{1}{N^n(s,a)}} \right]$$

$$\leq 4H\sqrt{S \ln(SAHN/\delta)} \left(\sum_{n,h} \sqrt{\frac{1}{N^n(s_h^n, a_h^n)}} + H \log(N/\delta) \right)$$

$$\leq 4H\sqrt{S \ln(SANH/\delta)} \left(2\sqrt{SAHN} + H \log(N/\delta) \right) \in \widetilde{O}\left(H^{1.5} S \sqrt{AN} \right)$$

5. Final Step

Remember we had two failure events for bounding transitions errors.

$$\text{Regret}_N = \sum_{n=1}^{N} \left(V_0^\star(s_0) - V_0^{\pi^n}(s_0) \right) \le 2H\sqrt{S \ln(SAHN/\delta)} \sum_{n=1}^{N} \sum_{h=0}^{H-1} \mathbb{E}_{s,a \sim d_h^{\pi^n}} \left[\sqrt{\frac{1}{N^n(s,a)}} \right]$$

$$\le 4H\sqrt{S \ln(SAHN/\delta)} \left(\sum_{n,h} \sqrt{\frac{1}{N^n(s_h^n, a_h^n)}} + H \log(N/\delta) \right)$$

w.p. $1-\delta$

$$\le 4H\sqrt{S \ln(SANH/\delta)} \left(2\sqrt{SAHN} + H \log(N/\delta) \right) \in \widetilde{O}\left(H^{1.5} S \sqrt{AN} \right)$$

$\sum_{i=1}^{m} \frac{1}{\sqrt{i}} \le 2\sqrt{m}$

$\forall$ sequences (s_h^n, a_h^n)

$$\sum_{n=1}^{N} \sum_{h=0}^{H-1} \frac{1}{\sqrt{N^n(s_h^n, a_h^n)}} = \sum_{s,a} \sum_{i=1}^{N^N(s,a)} \frac{1}{\sqrt{i}} \le 2 \sum_{s,a} \sqrt{N^N(s,a)} \le 2\sqrt{SA \sum_{s,a} N^N(s,a)} \le 2\sqrt{SANH}$$

CS

if we define $N^n(s,a)$ = visits to (s,a)

High-level Idea: Exploration or Exploitation Tradeoff

Upper bound per-episode regret: $V_0^\star(s_0) - V_0^{\pi^n}(s_0) \leq \widehat{V}_0^n(s_0) - V_0^{\pi^n}(s_0)$

1. What if $\widehat{V}_0^n(s_0) - V_0^{\pi^n}(s_0) \leq \epsilon$?

Then π^n is close to $\pi^\star$, i.e., we are doing exploitation

for any algorithm e.g. UCBVI

$\exists$ MDP s.t. Regret $\geq \Omega(H\sqrt{SAK})$

High-level Idea: Exploration or Exploitation Tradeoff

Upper bound per-episode regret: $V_0^\star(s_0) - V_0^{\pi^n}(s_0) \leq \widehat{V}_0^n(s_0) - V_0^{\pi^n}(s_0)$

1. What if $\widehat{V}_0^n(s_0) - V_0^{\pi^n}(s_0) \leq \epsilon$?

Then π^n is close to $\pi^\star$, i.e., we are doing exploitation

2. What if $\widehat{V}_0^n(s_0) - V_0^{\pi^n}(s_0) \geq \epsilon$?

High-level Idea: Exploration or Exploitation Tradeoff

Upper bound per-episode regret: $V_0^\star(s_0) - V_0^{\pi^n}(s_0) \leq \widehat{V}_0^n(s_0) - V_0^{\pi^n}(s_0)$

1) optimism

2) simulation lemma

1. What if $\widehat{V}_0^n(s_0) - V_0^{\pi^n}(s_0) \leq \epsilon$?

Then π^n is close to $\pi^\star$, i.e., we are doing exploitation

2. What if $\widehat{V}_0^n(s_0) - V_0^{\pi^n}(s_0) \geq \epsilon$?

$$\epsilon \leq \widehat{V}_0^n(s_0) - V_0^{\pi^n}(s_0) \leq \sum_{h=0}^{H-1} \mathbb{E}_{s,a \sim d_h^{\pi^n}} \left[b_h^n(s,a) + (\widehat{P}^n(\cdot \mid s,a) - P(\cdot \mid s,a)) \cdot \widehat{V}_{h+1}^n \right]$$

High-level Idea: Exploration or Exploitation Tradeoff

Upper bound per-episode regret: $V_0^\star(s_0) - V_0^{\pi^n}(s_0) \leq \widehat{V}_0^n(s_0) - V_0^{\pi^n}(s_0)$

1. What if $\widehat{V}_0^n(s_0) - V_0^{\pi^n}(s_0) \leq \epsilon$?

Then π^n is close to $\pi^\star$, i.e., we are doing exploitation

2. What if $\widehat{V}_0^n(s_0) - V_0^{\pi^n}(s_0) \geq \epsilon$?

$$\epsilon \leq \widehat{V}_0^n(s_0) - V_0^{\pi^n}(s_0) \leq \sum_{h=0}^{H-1} \mathbb{E}_{s,a\sim d_h^{\pi^n}} \left[b_h^n(s,a) + (\widehat{P}^n(\cdot \mid s,a) - P(\cdot \mid s,a)) \cdot \widehat{V}_{h+1}^n \right]$$

We collect data at steps where bonus is large or model is wrong, i.e., exploration