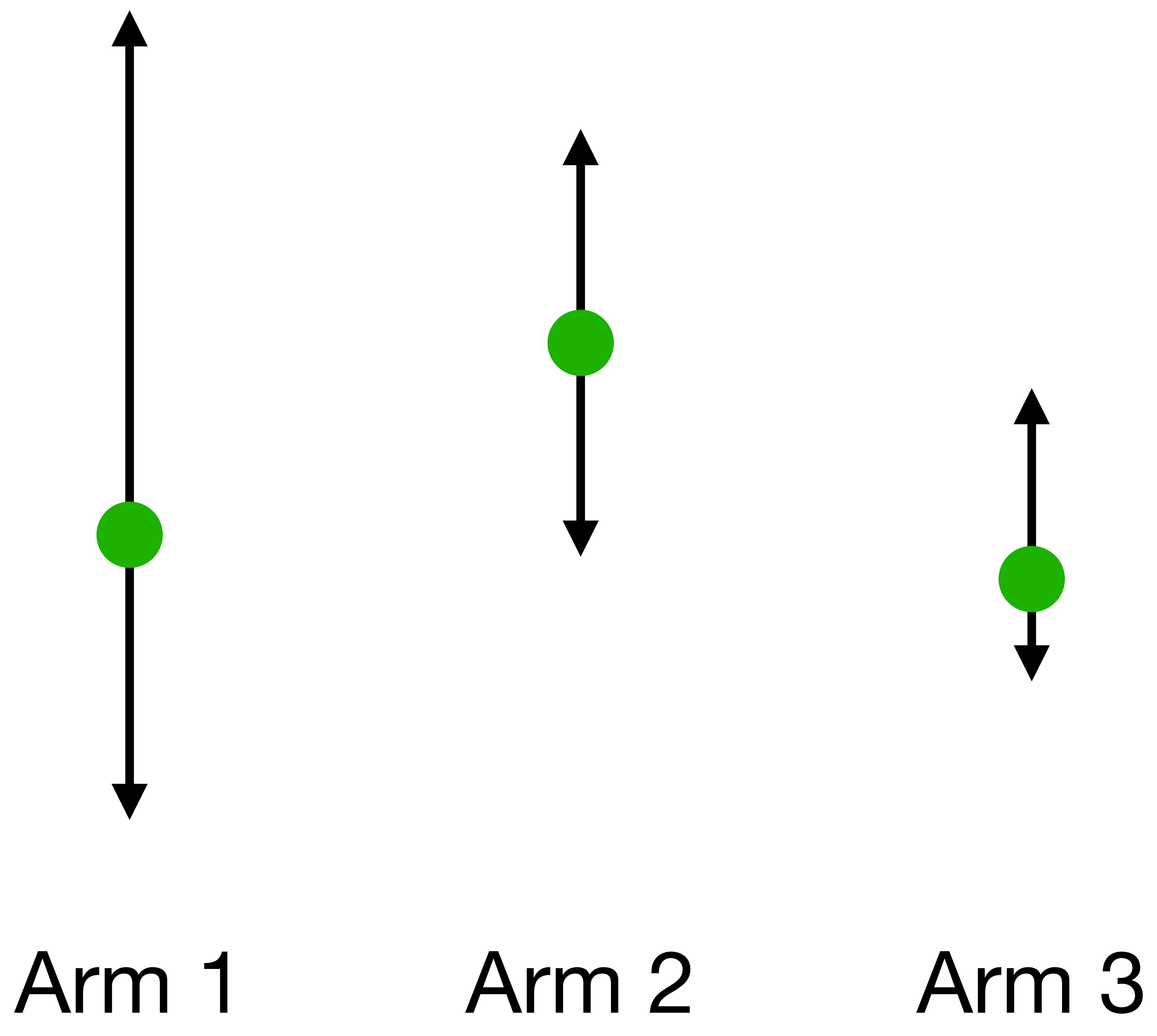
Exploration in MAB and Tabular MDPs

CS 6789: Foundations of Reinforcement Learning

Recap:

Multi-armed Bandits and UCB Algorithm



Let's formalize the intuition

Denote the optimal arm $I^\star = \arg\max_{i \in [K]} \mu_i$; recall $I_t = \arg\max_{i \in [K]} \hat{\mu}_t(i) + \sqrt{\frac{\ln(KT/\delta)}{N_t(i)}}$

Let's formalize the intuition

Denote the optimal arm $I^\star = \arg\max_{i \in [K]} \mu_i$; recall $I_t = \arg\max_{i \in [K]} \hat{\mu}_t(i) + \sqrt{\frac{\ln(KT/\delta)}{N_t(i)}}$

Regret-at-t $= \mu^\star - \mu_{I_t}$

Let's formalize the intuition

Denote the optimal arm $I^\star = \arg\max_{i\in[K]} \mu_i$; recall $I_t = \arg\max_{i\in[K]} \hat{\mu}_t(i) + \sqrt{\frac{\ln(KT/\delta)}{N_t(i)}}$

$$\text{Regret-at-t} = \mu^\star - \mu_{I_t}$$

$$\leq \widehat{\mu}_t(I_t) + \sqrt{\frac{\ln(TK/\delta)}{N_t(I_t)}} - \mu_{I_t}$$

Let's formalize the intuition

Denote the optimal arm $I^\star = \arg\max_{i \in [K]} \mu_i$; recall $I_t = \arg\max_{i \in [K]} \hat{\mu}_t(i) + \sqrt{\frac{\ln(KT/\delta)}{N_t(i)}}$

$$\text{Regret-at-t} = \mu^\star - \mu_{I_t}$$

Q: why?

$$\leq \widehat{\mu}_t(I_t) + \sqrt{\frac{\ln(TK/\delta)}{N_t(I_t)}} - \mu_{I_t}$$

Let's formalize the intuition

Denote the optimal arm $I^\star = \arg\max_{i \in [K]} \mu_i$; recall $I_t = \arg\max_{i \in [K]} \hat{\mu}_t(i) + \sqrt{\frac{\ln(KT/\delta)}{N_t(i)}}$

$$\text{Regret-at-t} = \mu^\star - \mu_{I_t}$$

Q: why?

$$\leq \hat{\mu}_t(I_t) + \sqrt{\frac{\ln(TK/\delta)}{N_t(I_t)}} - \mu_{I_t}$$

$$\leq 2\sqrt{\frac{\ln(TK/\delta)}{N_t(I_t)}}$$

Let's formalize the intuition

Denote the optimal arm $I^\star = \arg\max_{i \in [K]} \mu_i$; recall $I_t = \arg\max_{i \in [K]} \hat{\mu}_t(i) + \sqrt{\frac{\ln(KT/\delta)}{N_t(i)}}$

Regret-at-t $= \mu^\star - \mu_{I_t}$

Q: why?

$$\leq \hat{\mu}_t(I_t) + \sqrt{\frac{\ln(TK/\delta)}{N_t(I_t)}} - \mu_{I_t}$$

$$\leq 2\sqrt{\frac{\ln(TK/\delta)}{N_t(I_t)}}$$

Case 1:$N_t(I_t)$ is small

(i.e., uncertainty about I_t is large);

We pay regret, BUT we **explore** here,

as we just tried I_t at iter t!

Let's formalize the intuition

Denote the optimal arm $I^\star = \arg\max_{i \in [K]} \mu_i$; recall $I_t = \arg\max_{i \in [K]} \hat{\mu}_t(i) + \sqrt{\frac{\ln(KT/\delta)}{N_t(i)}}$

$$\text{Regret-at-t} = \mu^\star - \mu_{I_t}$$

$$\leq \hat{\mu}_t(I_t) + \sqrt{\frac{\ln(TK/\delta)}{N_t(I_t)}} - \mu_{I_t}$$

$$\leq 2\sqrt{\frac{\ln(TK/\delta)}{N_t(I_t)}}$$

Case 2: $N_t(I_t)$ is large, i.e., conf-interval of I_t is small,

Then we **exploit** here, as I_t is pretty good (the gap between $\mu^\star$ & μ_{I_t} is small)!

Let's formalize the intuition

Finally, let's add all per-iter regret together:

$$\text{Regret}_T = \sum_{t=0}^{T-1} \left(\mu^\star - \mu_{I_t} \right)$$

$$\leq \sum_{t=0}^{T-1} 2\sqrt{\frac{\ln(TK/\delta)}{N_t(I_t)}}$$

$$\leq 2\sqrt{\ln(TK/\delta)} \cdot \sum_{t=0}^{T-1} \sqrt{\frac{1}{N_t(I_t)}}$$

Let's formalize the intuition

Finally, let's add all per-iter regret together:

$$\text{Regret}_T = \sum_{t=0}^{T-1} \left(\mu^\star - \mu_{I_t} \right)$$

$$\leq \sum_{t=0}^{T-1} 2\sqrt{\frac{\ln(TK/\delta)}{N_t(I_t)}}$$

$$\leq 2\sqrt{\ln(TK/\delta)} \cdot \sum_{t=0}^{T-1} \sqrt{\frac{1}{N_t(I_t)}}$$

Lemma: $\sum_{t=0}^{T-1} \sqrt{\frac{1}{N_t(I_t)}} \leq O\left(\sqrt{KT}\right)$

Today: Efficient Learning in Finite Horizon tabular MDPs

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

Today: Efficient Learning in Finite Horizon tabular MDPs

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

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)

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:

$\{s_h^n, a_h^n, r_h^n\}_{h=0}^{H-1}$, 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:

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

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

Performance measure: REGRET

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

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

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

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?

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

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 $O(\sqrt{A^{SH}K})$

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

UCBVI: Optimistic Model-based Learning

Inside iteration n :

UCBVI: Optimistic Model-based Learning

Inside iteration n :

Use all previous data to estimate transitions $\widehat{P}_1^n, \ldots, \widehat{P}_{H-1}^n$

UCBVI: Optimistic Model-based Learning

Inside iteration n :

Use all previous data to estimate transitions $\widehat{P}_1^n, \ldots, \widehat{P}_{H-1}^n$

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 $\widehat{P}_1^n, \ldots, \widehat{P}_{H-1}^n$

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

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

UCBVI: Optimistic Model-based Learning

Inside iteration n :

Use all previous data to estimate transitions $\widehat{P}_1^n, \ldots, \widehat{P}_{H-1}^n$

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

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

Collect a new trajectory by executing π^n in the real world $\{P_h\}_{h=0}^{H-1}$ starting from s_0

UCBVI—Part 1: Model Estimation

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

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

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:

$$N_h^n(s,a) = \sum_{i=1}^{n-1} \mathbf{1}\{(s_h^i, a_h^i) = (s,a)\}, \forall s,a,h, \quad N_h^n(s,a,s') = \sum_{i=1}^{n-1} \mathbf{1}\{(s_h^i, a_h^i, s_{h+1}^i) = (s,a,s')\}, \forall s,a,h$$

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:

$$N_h^n(s,a) = \sum_{i=1}^{n-1} \mathbf{1}\{(s_h^i, a_h^i) = (s,a)\}, \forall s, a, h, \quad N_h^n(s,a,s') = \sum_{i=1}^{n-1} \mathbf{1}\{(s_h^i, a_h^i, s_{h+1}^i) = (s,a,s')\}, \forall s, a, h$$

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

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

UCBVI—Part 2: Reward Bonus Design and Value Iteration

Let us consider the very beginning of episode n:

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

UCBVI—Part 2: Reward Bonus Design and Value Iteration

Let us consider the very beginning of episode n:

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

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

UCBVI—Part 2: Reward Bonus Design and Value Iteration

Let us consider the very beginning of episode n:

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

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

Encourage to explore new state-actions

UCBVI—Part 2: Reward Bonus Design and Value Iteration

Let us consider the very beginning of episode n:

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

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

Encourage to explore new state-actions

Value Iteration (aka DP) at episode n using $\{\widehat{P}_h^n\}_h$ **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:

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

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

Encourage to explore new state-actions

Value Iteration (aka DP) at episode n using $\{\widehat{P}_h^n\}_h$ **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:

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

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

Encourage to explore new state-actions

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

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

UCBVI—Part 2: Reward Bonus Design and Value Iteration

Let us consider the very beginning of episode n:

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

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

Encourage to explore new state-actions

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

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

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

UCBVI—Part 2: Reward Bonus Design and Value Iteration

Let us consider the very beginning of episode n:

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

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

Encourage to explore new state-actions

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

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

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

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

UCBVI: Put All Together

For $n = 1 \to N$:

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

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

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

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

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

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

$$\mathbb{E}\left[\text{Regret}_N\right] := \mathbb{E}\left[\sum_{n=1}^{N}\left(V^\star - V^{\pi^n}\right)\right] \leq \widetilde{O}\left(H^2\sqrt{S^2AN}\right)$$

Theorem: UCBVI Regret Bound

$$\mathbb{E}\left[\text{Regret}_N\right] := \mathbb{E}\left[\sum_{n=1}^{N}\left(V^\star - V^{\pi^n}\right)\right] \leq \widetilde{O}\left(H^2\sqrt{S^2AN}\right)$$

Remarks:

Note that we consider expected regret here (policy π^n is a random quantity).
High probability version is not hard to get (need to do a martingale argument)

Theorem: UCBVI Regret Bound

$$\mathbb{E}\left[\text{Regret}_N\right] := \mathbb{E}\left[\sum_{n=1}^{N}\left(V^\star - V^{\pi^n}\right)\right] \leq \widetilde{O}\left(H^2\sqrt{S^2AN}\right)$$

Remarks:

Note that we consider expected regret here (policy π^n is a random quantity).
High probability version is not hard to get (need to do a martingale argument)

Dependency on H and S are suboptimal; but the **same** algorithm can achieve $H^2\sqrt{SAN}$ 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 $\left(\left(\widehat{P}_h^n(\cdot \mid s,a) - P_h(\cdot \mid s,a)\right) \cdot V_{h+1}^\star\right)$

Outline of Proof

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

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

Outline of Proof

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

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

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

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

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

$$\widehat{P}_h^n(s' \mid s, a) = \frac{N_h^n(s, a, s')}{N_h^n(s, a)}, \forall h, s, a, s'$$

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

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

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

$$\widehat{P}_h^n(s' \mid s, a) = \frac{N_h^n(s, a, s')}{N_h^n(s, a)}, \forall h, s, a, s'$$

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

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

Bonus $b_h^n(s, a)$

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

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

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

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

Bonus $b_h^n(s,a)$

From now on, assume this event being true

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

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

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

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

Bonus $b_h^n(s,a)$

From now on, assume this event being true

Intuition:

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

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

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

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

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 $\mathbb{E}_{s' \sim P_h(\cdot|s,a)} f(s')$

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

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

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

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

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

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:

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

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

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:

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

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

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:

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

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

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

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

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:

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

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

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

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

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

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:

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

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

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

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

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

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

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:

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

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

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

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

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

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

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

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]:

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

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]:

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

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

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]:

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

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

$$\leq r_0(s_0, \pi^n(s_0)) + b^n_h(s_0, \pi^n(s_0)) + \widehat{P}^n_0(\cdot \mid s_0, \pi^n(s_0)) \cdot \widehat{V}^n_1 - r_0(s_0, \pi^n(s_0)) - P_0(\cdot \mid s_0, \pi^n(s_0)) \cdot V^{\pi^n}_1$$

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]:

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

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

$$\leq r_0(s_0, \pi^n(s_0)) + b^n_h(s_0, \pi^n(s_0)) + \widehat{P}^n_0(\cdot \mid s_0, \pi^n(s_0)) \cdot \widehat{V}^n_1 - r_0(s_0, \pi^n(s_0)) - P_0(\cdot \mid s_0, \pi^n(s_0)) \cdot V^{\pi^n}_1$$

$$= b^n_h(s_0, \pi^n(s_0)) + \widehat{P}^n_0(\cdot \mid s_0, \pi^n(s_0)) \cdot \widehat{V}^n_1 - P_0(\cdot \mid s_0, \pi^n(s_0)) \cdot V^{\pi^n}_1$$

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]:

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

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

$$\leq r_0(s_0, \pi^n(s_0)) + b^n_h(s_0, \pi^n(s_0)) + \widehat{P}^n_0(\cdot \mid s_0, \pi^n(s_0)) \cdot \widehat{V}^n_1 - r_0(s_0, \pi^n(s_0)) - P_0(\cdot \mid s_0, \pi^n(s_0)) \cdot V^{\pi^n}_1$$

$$= b^n_h(s_0, \pi^n(s_0)) + \widehat{P}^n_0(\cdot \mid s_0, \pi^n(s_0)) \cdot \widehat{V}^n_1 - P_0(\cdot \mid s_0, \pi^n(s_0)) \cdot V^{\pi^n}_1$$

$$= b^n_h(s_0, \pi^n(s_0)) + \left(\widehat{P}^n_0(\cdot \mid s_0, \pi^n(s_0)) - P_0(\cdot \mid s_0, \pi^n(s_0)) \right) \cdot \widehat{V}^n_1 + P_0(\cdot \mid s_0, \pi^n(s_0)) \cdot \left(\widehat{V}^n_1 - V^{\pi^n}_1 \right)$$

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]:

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

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

$$\leq r_0(s_0,\pi^n(s_0)) + b^n_h(s_0,\pi^n(s_0)) + \widehat{P}^n_0(\cdot \mid s_0,\pi^n(s_0))\cdot \widehat{V}^n_1 - r_0(s_0,\pi^n(s_0)) - P_0(\cdot \mid s_0,\pi^n(s_0))\cdot V^{\pi^n}_1$$

$$= b^n_h(s_0,\pi^n(s_0)) + \widehat{P}^n_0(\cdot \mid s_0,\pi^n(s_0))\cdot \widehat{V}^n_1 - P_0(\cdot \mid s_0,\pi^n(s_0))\cdot V^{\pi^n}_1$$

$$= b^n_h(s_0,\pi^n(s_0)) + \left(\widehat{P}^n_0(\cdot \mid s_0,\pi^n(s_0)) - P_0(\cdot \mid s_0,\pi^n(s_0)) \right)\cdot \widehat{V}^n_1 + P_0(\cdot \mid s_0,\pi^n(s_0))\cdot \left(\widehat{V}^n_1 - V^{\pi^n}_1 \right)$$

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

4. Upper bounding Regret via Simulation Lemma

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

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

4. Upper bounding Regret via Simulation Lemma

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

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

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

Let's do Holder's inequality

4. Upper bounding Regret via Simulation Lemma

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

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

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 $V_h^\star$) !!!

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

Let's do Holder's inequality

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

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

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 $V_h^\star$) !!!

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

Let's do Holder's inequality

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

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

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

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

Let's do Holder's inequality

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

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

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

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

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 $V_h^\star$) !!!

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

Let's do Holder's inequality

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

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

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

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

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.

$$\mathbb{E}\left[\text{Regret}_N\right] = \mathbb{E}\left[\mathbf{1}\{\text{events hold}\}\sum_{n=1}^{N}\left(V_0^{\star}(s_0) - V_0^{\pi^n}(s_0)\right)\right] + \mathbb{E}\left[\mathbf{1}\{\text{events don't hold}\}\sum_{n=1}^{N}\left(V_0^{\star}(s_0) - V_0^{\pi^n}(s_0)\right)\right]$$

5. Final Step

Remember we had two failure events for bounding transitions errors.

$$\mathbb{E}\left[\text{Regret}_N\right] = \mathbb{E}\left[\mathbf{1}\{\text{events hold}\}\sum_{n=1}^{N}\left(V_0^\star(s_0) - V_0^{\pi^n}(s_0)\right)\right] + \mathbb{E}\left[\mathbf{1}\{\text{events don't hold}\}\sum_{n=1}^{N}\left(V_0^\star(s_0) - V_0^{\pi^n}(s_0)\right)\right]$$

$$\leq \mathbb{E}\left[\mathbf{1}\{\text{events hold}\}\sum_{n=1}^{N}\left(V_0^\star(s_0) - V_0^{\pi^n}(s_0)\right)\right] + \mathbb{P}(\text{events don't hold}) \cdot NH$$

5. Final Step

Remember we had two failure events for bounding transitions errors.

$$\mathbb{E}\left[\text{Regret}_N\right] = \mathbb{E}\left[\mathbf{1}\{\text{events hold}\}\sum_{n=1}^{N}\left(V_0^\star(s_0) - V_0^{\pi^n}(s_0)\right)\right] + \mathbb{E}\left[\mathbf{1}\{\text{events don't hold}\}\sum_{n=1}^{N}\left(V_0^\star(s_0) - V_0^{\pi^n}(s_0)\right)\right]$$

$$\leq \mathbb{E}\left[\mathbf{1}\{\text{events hold}\}\sum_{n=1}^{N}\left(V_0^\star(s_0) - V_0^{\pi^n}(s_0)\right)\right] + \mathbb{P}(\text{events don't hold}) \cdot NH$$

$$\leq H\sqrt{S\ln(SANH/\delta)}\,\mathbb{E}\left[\sum_{n=1}^{N}\sum_{h=0}^{H-1}\frac{1}{\sqrt{N_h^n(s_h^n, a_h^h)}}\right] + 2\delta NH$$

5. Final Step

$$\sum_{n=1}^{N} \sum_{h=0}^{H-1} \frac{1}{\sqrt{N_h^n(s_h^n, a_h^n)}}$$

5. Final Step

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

5. Final Step

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

5. Final Step

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

5. Final Step

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

$$\leq \sum_{h=0}^{H-1}\sqrt{SAN} = H\sqrt{SAN}$$

5. Final Step

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

$$\leq \sum_{h=0}^{H-1}\sqrt{SAN} = H\sqrt{SAN}$$

$$\mathbb{E}\left[\text{Regret}_N\right] \leq 2H^2S\sqrt{AN\ln(SAHN/\delta)} + 2\delta NH$$

5. Final Step

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

$$\le \sum_{h=0}^{H-1}\sqrt{SAN} = H\sqrt{SAN}$$

$$\mathbb{E}\left[\text{Regret}_N\right] \le 2H^2S\sqrt{AN\ln(SAHN/\delta)} + 2\delta NH$$

Set $\delta = 1/(HN)$

$$\le 2H^2S\sqrt{AN\cdot\ln(SAH^2N^2)} = \widetilde{O}\left(H^2S\sqrt{AN}\right)$$

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

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

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

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

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

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