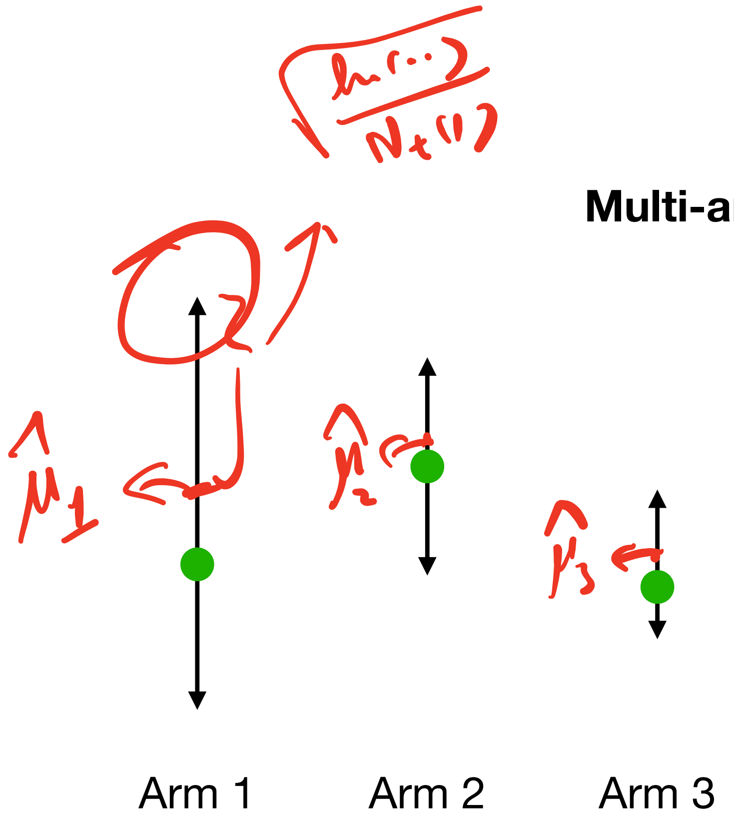
Exploration in MAB and Tabular MDPs

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 $I_t = \arg\max_{i\in[K]} \underbrace{\hat{\mu}_t(i) + \sqrt{\frac{\ln(KT/\delta)}{N_t(i)}}}_{\text{UCB of } i}$

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 $I_t = \arg\max_{i \in [K]} \hat{\mu}_t(i) + \sqrt{\frac{\ln(KT/\delta)}{N_t(i)}}$

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

Q: why?

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

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?

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

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

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?

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

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

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 $I_t = \arg\max_{i \in [K]} \hat{\mu}_t(i) + \sqrt{\frac{\ln(KT/\delta)}{N_t(i)}}$

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

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

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

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:

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

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

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

$1 + \sqrt{\frac{1}{2}} + \sqrt{\frac{1}{3}} + \sqrt{\frac{1}{4}} \cdots \sqrt{\frac{1}{T}}$

$\lesssim \sqrt{T}$

Let's formalize the intuition

a_1

$t=0 \quad 1 \quad 2 \quad 3 \quad \cdots \quad T$

$\sqrt{\frac{1}{1}} \quad \sqrt{\frac{1}{2}} \quad \sqrt{\frac{1}{3}} \quad \sqrt{\frac{1}{N_t(a_1)}}$

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

$$\sum_{i=1}^{K}\sum_{t=0}^{T-1} \mathbb{1}(I_t = i)\sqrt{\frac{1}{N_t(i)}}$$

$$= \sum_{i=1}^{K}\sum_{j=1}^{N_T(i)}\sqrt{\frac{1}{j}} \leq \sum_{i=1}^{K}\sqrt{N_T(i)}$$

Today: Efficient Learning in Finite Horizon tabular MDPs

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

$$\sum_{i=1}^{K} \sqrt{N_T(i)} \le \sqrt{K \Big(\underbrace{\sum_{i=1}^{K} N_T(i)}_{T} \Big)} = \sqrt{KT}$$

$$|a^T b| \le \sqrt{\|a\|_2^2 \|b\|_2^2}$$

Today: Efficient Learning in Finite Horizon tabular MDPs

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

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

$$s_0 \xrightarrow{a_0 \sim \pi^n(s_0)} s_1 \xrightarrow{a_1 \sim \pi^n} s_2 \rightarrow \cdots \quad s_{H-1}$$

$$r_0 \quad r_1 \quad r_2 \quad r_{H-1}$$

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

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

R(s)

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

$d_h^\pi(s,a)$: 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

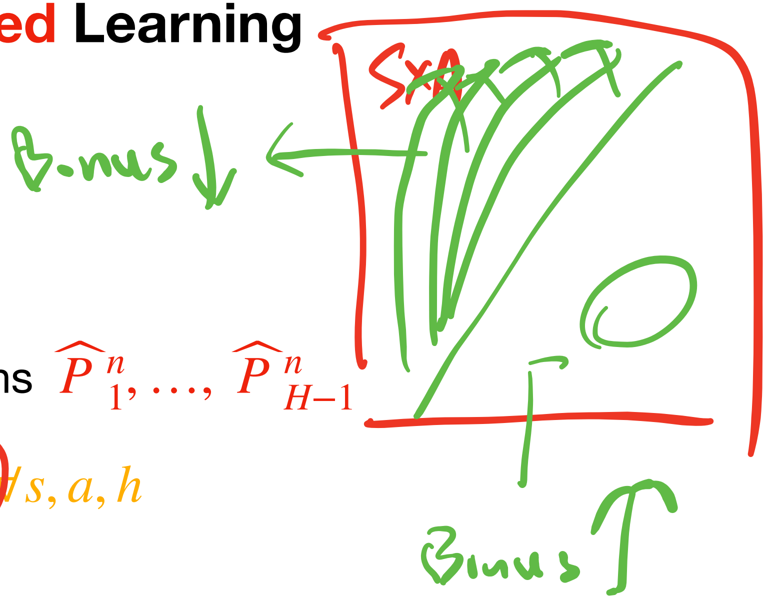
Inside iteration n :

UCBVI: Optimistic Model-based Learning

Inside iteration n :

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

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 $\widehat{P}^n_1, \ldots, \widehat{P}^n_{H-1}$

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

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

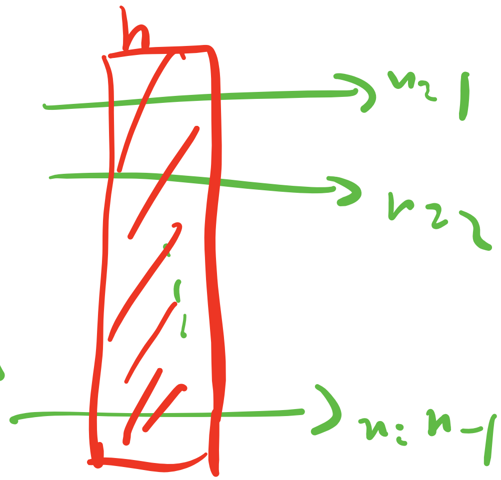
Collect a new trajectory by executing π^n in the real world $\{P_h\}_{h=0}^{H-1}$ starting from s_0

UCBVI—Part 1: Model Estimation

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

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

history data



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:

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:

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

Let's also maintain some statistics using these datasets:

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

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

$$\widehat{P}_h^n(s'|s,a) = \frac{N_h^n(s,a,s')}{N_h^n(s,a)} \qquad \frac{N_h^n(s,a,s')}{\max\{1, N_h^n(s,a)\}}$$

UCBVI—Part 2: Reward Bonus Design and Value Iteration

Let us consider the very beginning of episode n:

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

UCBVI—Part 2: Reward Bonus Design and Value Iteration

Let us consider the very beginning of episode n:

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

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

UCBVI—Part 2: Reward Bonus Design and Value Iteration

Let us consider the very beginning of episode n:

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

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

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:

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

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

Encourage to explore new state-actions

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

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

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 $\{r_h + b_h^n\}_h$

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

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** $\{r_h + b_h^n\}_h$

$\in [0, H]$

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

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

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

UCBVI: Put All Together

For $n = 1 \rightarrow N$:

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

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., $\widehat{V}_h^n(s) \geq V_h^\star(s), \forall h, n, s, a$

optimism

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

Regret at n

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: $\underbrace{\widehat{V}_0^n(s_0)}_{(\hat{P},\, r+b)} - \underbrace{V^{\pi^n}(s_0)}_{(P,\, r)}$

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

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

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 $1-\delta$:

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

union Bound

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 $b_h^n(s, a)$

$V^\star$