

Exploration in Tabular MDPs

Sham Kakade and Kianté Brantley

CS 2824: Foundations of Reinforcement Learning

Announcements

1. HW2: Out 02/24, Due 03/13

2. Course Project Website

<https://harvard-cs2824-s26.github.io/CS2824projects.html>

Recap: Explore and Exploration Tradeoff

Case 1: I_t has large conf-interval, which means that it has not been tried many times yet (high uncertainty)

Thus, we do exploration in this case!

Recap: Explore and Exploration Tradeoff

Case 1: I_t has large conf-interval, which means that it has not been tried many times yet (high uncertainty)

Thus, we do exploration in this case!

Case 2: I_t has small conf-interval, then it is simply a good arm, i.e., its true mean is pretty high!

Thus, we do exploitation in this case!

Let's formalize the intuition

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

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

Let's formalize the intuition

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

$$\begin{aligned} \text{Regret-at-t} &= \mu^* - \mu_{I_t} \\ &\leq \hat{\mu}_t(I_t) + \sqrt{\frac{\ln(TK/\delta)}{N_t(I_t)}} - \mu_{I_t} \end{aligned}$$

Let's formalize the intuition

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

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

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

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

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

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

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

$$\begin{aligned} \text{Regret-at-t} &= \mu^* - \mu_{I_t} \\ &\leq \hat{\mu}_t(I_t) + \sqrt{\frac{\ln(TK/\delta)}{N_t(I_t)}} - \mu_{I_t} \\ &\leq 2\sqrt{\frac{\ln(TK/\delta)}{N_t(I_t)}} \end{aligned}$$

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 μ^* & μ_{I_t} is small)!

Let's formalize the intuition

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

$$\begin{aligned}\text{Regret}_T &= \sum_{t=0}^{T-1} \left(\mu^\star - \mu_{I_t} \right) \\ &\leq \sum_{t=0}^{T-1} 2 \sqrt{\frac{\ln(TK/\delta)}{N_t(I_t)}} \\ &\leq 2 \sqrt{\ln(TK/\delta)} \cdot \sum_{t=0}^{T-1} \sqrt{\frac{1}{N_t(I_t)}}\end{aligned}$$

Let's formalize the intuition

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

$$\begin{aligned} \text{Regret}_T &= \sum_{t=0}^{T-1} \left(\mu^\star - \mu_{I_t} \right) \\ &\leq \sum_{t=0}^{T-1} 2 \sqrt{\frac{\ln(TK/\delta)}{N_t(I_t)}} \\ &\leq 2 \sqrt{\ln(TK/\delta)} \cdot \sum_{t=0}^{T-1} \sqrt{\frac{1}{N_t(I_t)}} \end{aligned}$$

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

UCB Regret:

[Theorem (informal)] With high probability, UCB has the following regret:

$$\text{Regret}_T = \tilde{O}\left(\sqrt{KT}\right)$$

Today: Efficient Learning in Finite Horizon tabular MDPs

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

Today: Efficient Learning in Finite Horizon tabular MDPs

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

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 $\mathcal{M} = \{ \{r_h\}_{h=0}^{H-1}, \{P_h\}_{h=0}^H, H, \mu, S, A \}$

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)

EXPLORATION!

Why we need strategic exploration?

Initialization: s_0



Thrun '92

Length of chain is H

Why we need strategic exploration?

Initialization: s_0



Thrun '92

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

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

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 :

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

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

Performance measure: REGRET

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

Notations for Today

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

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

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

UCBVI: Optimistic Model-based Learning

Inside iteration n :

UCBVI: Optimistic Model-based Learning

Inside iteration n :

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

UCBVI: Optimistic Model-based Learning

Inside iteration n :

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

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 $\hat{P}_1^n, \dots, \hat{P}_{H-1}^n$

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

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

UCBVI: Optimistic Model-based Learning

Inside iteration n :

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

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

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

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

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

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 :

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

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

UCBVI – Part 2: Reward Bonus Design and Value Iteration

Let us consider the very beginning of episode n :

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

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

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 :

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

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

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

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

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

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

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

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

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

$$\widehat{V}_h^n(s) = \max_a \widehat{Q}_h^n(s, a), \quad \pi_h^n(s) = \arg \max_a \widehat{Q}_h^n(s, a), \forall s \quad \left\| \widehat{V}_h^n \right\|_{\infty} \leq H, \forall h, n$$

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

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

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

Theorem: UCBVI Regret Bound

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

Theorem: UCBVI Regret Bound

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

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

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

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]

Outline of Proof

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

Outline of Proof

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

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

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

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

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

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 π^\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 π^\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 π^\star , i.e., we are doing exploitation

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

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

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 π^\star , i.e., we are doing exploitation

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

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

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

1. Model Error using Hoeffding'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 Hoeffding'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 $1 - \delta$:

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

1. Model Error using Hoeffding'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 $1 - \delta$:

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

Bonus $b_h^n(s, a)$

1. Model Error using Hoeffding'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 $1 - \delta$:

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

Bonus $b_h^n(s, a)$

From now on, assume this event being true

1. Model Error using Hoeffding'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 $1 - \delta$:

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

Bonus $b_h^n(s, a)$

From now on, assume this event being true

Intuition:

1. Model Error using Hoeffding'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 $1 - \delta$:

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

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 Hoeffding'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 $1 - \delta$:

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

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 :

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

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:

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

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

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:

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

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

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

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

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:

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

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

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

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

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:

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

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

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

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

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

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

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:

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

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

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

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

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

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

$$\geq b_h^n(s, a) - b_h^n(s, a) = 0, \quad \forall s, a$$

3. Upper Bounding Regret using Optimism

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

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

Recall simulation lemma — the lemma measures the difference of a policy under two MDPs

4. Upper bounding Regret via Simulation Lemma

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

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

Lemma [Simulation lemma]:

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

4. Upper bounding Regret via Simulation Lemma

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

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

Lemma [Simulation lemma]:

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

$$\widehat{V}_0^n(s_0) - V_0^{\pi^n}(s_0) = \widehat{Q}_0^n(s_0, \pi^n(s_0)) - Q_0^{\pi^n}(s_0, \pi^n(s_0))$$

4. Upper bounding Regret via Simulation Lemma

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

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

Lemma [Simulation lemma]:

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

$$\widehat{V}_0^n(s_0) - V_0^{\pi^n}(s_0) = \widehat{Q}_0^n(s_0, \pi^n(s_0)) - Q_0^{\pi^n}(s_0, \pi^n(s_0))$$

$$\leq r_0(s_0, \pi^n(s_0)) + b_h^n(s_0, \pi^n(s_0)) + \widehat{P}_0^n(\cdot | s_0, \pi^n(s_0)) \cdot \widehat{V}_1^n - r_0(s_0, \pi^n(s_0)) - P_0(\cdot | s_0, \pi^n(s_0)) \cdot V_1^{\pi^n}$$

4. Upper bounding Regret via Simulation Lemma

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

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

Lemma [Simulation lemma]:

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

$$\widehat{V}_0^n(s_0) - V_0^{\pi^n}(s_0) = \widehat{Q}_0^n(s_0, \pi^n(s_0)) - Q_0^{\pi^n}(s_0, \pi^n(s_0))$$

$$\leq r_0(s_0, \pi^n(s_0)) + b_h^n(s_0, \pi^n(s_0)) + \widehat{P}_0^n(\cdot | s_0, \pi^n(s_0)) \cdot \widehat{V}_1^n - r_0(s_0, \pi^n(s_0)) - P_0(\cdot | s_0, \pi^n(s_0)) \cdot V_1^{\pi^n}$$

$$= b_h^n(s_0, \pi^n(s_0)) + \widehat{P}_0^n(\cdot | s_0, \pi^n(s_0)) \cdot \widehat{V}_1^n - P_0(\cdot | s_0, \pi^n(s_0)) \cdot V_1^{\pi^n}$$

4. Upper bounding Regret via Simulation Lemma

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

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

Lemma [Simulation lemma]:

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

$$\widehat{V}_0^n(s_0) - V_0^{\pi^n}(s_0) = \widehat{Q}_0^n(s_0, \pi^n(s_0)) - Q_0^{\pi^n}(s_0, \pi^n(s_0))$$

$$\leq r_0(s_0, \pi^n(s_0)) + b_h^n(s_0, \pi^n(s_0)) + \widehat{P}_0^n(\cdot | s_0, \pi^n(s_0)) \cdot \widehat{V}_1^n - r_0(s_0, \pi^n(s_0)) - P_0(\cdot | s_0, \pi^n(s_0)) \cdot V_1^{\pi^n}$$

$$= b_h^n(s_0, \pi^n(s_0)) + \widehat{P}_0^n(\cdot | s_0, \pi^n(s_0)) \cdot \widehat{V}_1^n - P_0(\cdot | s_0, \pi^n(s_0)) \cdot V_1^{\pi^n}$$

$$= b_h^n(s_0, \pi^n(s_0)) + \left(\widehat{P}_0^n(\cdot | s_0, \pi^n(s_0)) - P_0(\cdot | s_0, \pi^n(s_0)) \right) \cdot \widehat{V}_1^n + P_0(\cdot | s_0, \pi^n(s_0)) \cdot \left(\widehat{V}_1^n - V_1^{\pi^n} \right)$$

4. Upper bounding Regret via Simulation Lemma

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

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

Lemma [Simulation lemma]:

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

$$\widehat{V}_0^n(s_0) - V_0^{\pi^n}(s_0) = \widehat{Q}_0^n(s_0, \pi^n(s_0)) - Q_0^{\pi^n}(s_0, \pi^n(s_0))$$

$$\leq r_0(s_0, \pi^n(s_0)) + b_h^n(s_0, \pi^n(s_0)) + \widehat{P}_0^n(\cdot | s_0, \pi^n(s_0)) \cdot \widehat{V}_1^n - r_0(s_0, \pi^n(s_0)) - P_0(\cdot | s_0, \pi^n(s_0)) \cdot V_1^{\pi^n}$$

$$= b_h^n(s_0, \pi^n(s_0)) + \widehat{P}_0^n(\cdot | s_0, \pi^n(s_0)) \cdot \widehat{V}_1^n - P_0(\cdot | s_0, \pi^n(s_0)) \cdot V_1^{\pi^n}$$

$$= b_h^n(s_0, \pi^n(s_0)) + \left(\widehat{P}_0^n(\cdot | s_0, \pi^n(s_0)) - P_0(\cdot | s_0, \pi^n(s_0)) \right) \cdot \widehat{V}_1^n + P_0(\cdot | s_0, \pi^n(s_0)) \cdot \left(\widehat{V}_1^n - V_1^{\pi^n} \right)$$

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

4. Upper bounding Regret via Simulation Lemma

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

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

4. Upper bounding Regret via Simulation Lemma

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

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

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

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

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

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

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

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

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

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

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

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

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.

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

5. Final Step

Remember we had two failure events for bounding transitions errors.

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

5. Final Step

Remember we had two failure events for bounding transitions errors.

$$\begin{aligned}
 \mathbb{E} [\text{Regret}_N] &= \mathbb{E} \left[\mathbf{1}\{\text{events hold}\} \sum_{n=1}^N (V_0^\star(s_0) - V_0^{\pi^n}(s_0)) \right] + \mathbb{E} \left[\mathbf{1}\{\text{events don't hold}\} \sum_{n=1}^N (V_0^\star(s_0) - V_0^{\pi^n}(s_0)) \right] \\
 &\leq \mathbb{E} \left[\mathbf{1}\{\text{events hold}\} \sum_{n=1}^N (V_0^\star(s_0) - V_0^{\pi^n}(s_0)) \right] + \mathbb{P}(\text{events don't hold}) \cdot NH \\
 &\leq H\sqrt{S \ln(SANH/\delta)} \mathbb{E} \left[\sum_{n=1}^N \sum_{h=0}^{H-1} \frac{1}{\sqrt{N_h^n(s_h^n, a_h^h)}} \right] + 2\delta NH
 \end{aligned}$$

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

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

5. Final Step

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

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

5. Final Step

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

$$\mathbb{E} [\text{Regret}_N] \leq 2H^2S\sqrt{AN \ln(SAHN/\delta)} + 2\delta NH \quad \text{Set } \delta = 1/(HN)$$

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

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 π^\star , i.e., we are doing exploitation

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

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

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