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

High-level Idea: Exploration or Exploitation Tradeoff

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

$\pi^n, \hat{P}$, etc

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 $\pi^{\star}$, 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

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

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

f: V^*

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

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

$O(H\sqrt{\frac{1}{N_h^n(s,a)}})$

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

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

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 $b_h^n(s, a)$

From now on, assume this event being true

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

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

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

$$= b_h^n(s,a) + \left(\widehat{P}_h^n(\cdot \mid s,a) - P_h(\cdot \mid s,a) \right) \cdot V_{h+1}^\star \leq 4\sqrt{\frac{1}{N_h^n(s,a)}} \approx b_h^n$$

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

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

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

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

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

$\widehat{Q}_{sa}^n - Q_{sa}^\star \geq 0$

$\widehat{Q}_{sa}^n \geq Q_{sa}^\star$ $\Rightarrow$ $\widehat{V}_{sa}^n \geq V_{sa}^\star$

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

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

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

Lemma [Simulation lemma]:

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

4. Upper bounding Regret via Simulation Lemma

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

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

Lemma [Simulation lemma]:

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

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

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]:

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

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

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

min(a,b) ≤ a

update

Hel equ.

4. Upper bounding Regret via Simulation Lemma

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

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

Lemma [Simulation lemma]:

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

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

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

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

+/−

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]:

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

stage

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

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

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

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

step 2

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]:

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

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

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

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

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

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

4. Upper bounding Regret via Simulation Lemma

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

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

$|\,(\widehat{P} - P^*)V^*\,| \le H\sqrt{\frac{1}{N_h^n(s,a)}}$

4. Upper bounding Regret via Simulation Lemma

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

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

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

Let's do Holder's inequality

4. Upper bounding Regret via Simulation Lemma

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

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

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

Let's do Holder's inequality

$\|a \cdot b\| \le \|a\|_f \cdot \|b\|_g$ $\forall t \in [0,1]$

$\|V\|_\infty \le H$

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

4. Upper bounding Regret via Simulation Lemma

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

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

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

Let's do Holder's inequality

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

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

4. Upper bounding Regret via Simulation Lemma

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

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

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

Let's do Holder's inequality

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

$b_h^n = H\sqrt{\frac{\ln(\cdots)}{N_h^n(s,a)}} \le H\sqrt{\frac{S\ln(\cdots)}{N_{\text{prev}}(s,a)}}$

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

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

4. Upper bounding Regret via Simulation Lemma

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

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

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

Let's do Holder's inequality

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

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

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

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

4. Upper bounding Regret via Simulation Lemma

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

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

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

Let's do Holder's inequality

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

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

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

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

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

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

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

5. Final Step

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

CA - inequality

$$\sum_{s,a}1\cdot\sqrt{N_h^n(s,a)} \leq \sqrt{\sum_{s,a}1}\cdot\sqrt{\sum_{s,a}N_h^n(s,a)}$$

$$\sum_{s,a}1 = SA$$

$$\sqrt{SA}$$

5. Final Step

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

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

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

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

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

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

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

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

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