Provably Efficient Reinforcement Learning with Aggregated States

Xiuwen Wang

August 13, 2021

Reference:
1. Shi Dong, Benjamin Van Roy, and Zhengyuan Zhou, (2020) Provably Efficient Reinforcement Learning with Aggregated States. arXiv:1912.06366
2. Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, Michael I. Jordan, (2018) Is Q-learning Provably Efficient. arXiv:1807.03765
3. Shi Dong, Benjamin Van Roy, Zhengyuan Zhou, (2021) Simple Agent, Complex Environment: Efficient Reinforcement Learning with Agent States. arXiv:2102.05261.

Overview

- RL with Aggregated States:
  1. Motivation
  2. Problem Formulation
  3. Aggregated Q-learning with Upper Confidence Bounds
  4. Main Results
- RL with Agent States

Introduction

- RL algorithms with tabular representations
  $\Rightarrow$ Data and learning time grow with the number of state-action pairs.
- How to address this problem?
  $\Rightarrow$ State aggregation.

  1. partition the set of all state-action pairs, each cell representing an aggregate state.
  2. learn the value function for each cell.
  3. $\tilde{O}\left(\sqrt{H^5MK} + \epsilon HK\right)$ worst-case regret bound without assumptions on the structure of the environment.

Problem Formulation and Notations

- finite state space $\mathcal{S}$ and action space $\mathcal{A}$ with cardinality S and A, respectively
- K episodes, each consists of H stages and produces a sequence

$$s_1, a_1, \ldots, s_H, a_H$$

- deterministic reward $R_h(s, a) \in [0, 1]$, system dynamics $\mathrm{P}_h^{s,a}(s')$
- $0 \leq V_h^{\pi} \leq V_h^{*} \leq H$
- $\mathrm{Regret}(K) = \sum_{k=1}^{K} V_1^{*}(s_1) - V_1^{\pi_k}(s_1)$

Problem Formulation and Notations

- the set of aggregate states $\Phi = [M]$
- $\phi_h : \mathcal{S} \times \mathcal{A} \mapsto \Phi$

.

Definition (ϵ-error aggregation)

$\{\phi_h\}_{h=1}^{H}$ is an ϵ-error aggregated state representation (or ϵ -error aggregation) of an MDP, if for all $s, s' \in \mathcal{S}, a, a' \in \mathcal{A}$ and $h \in [H]$ such that $\phi_h(s, a) = \phi_h(s', a')$,

$$|Q_h^*(s, a) - Q_h^*(s', a')| \leq \epsilon$$

Q-learning with Upper Confidence Bounds

Algorithm 1 Q-learning with UCB-Hoeffding

initialize $Q_h(x,a) \leftarrow H$ and $N_h(x,a) \leftarrow 0$ for all $(x,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$.
for episode $k = 1, \ldots, K$ **do**
  receive x_1.
  for step $h = 1, \ldots, H$ **do**
    Take action $a_h \leftarrow \operatorname{argmax}_{a'} Q_h(x_h, a')$, and observe x_{h+1}.
    $t = N_h(x_h, a_h) \leftarrow N_h(x_h, a_h) + 1;\ \ b_t \leftarrow c\sqrt{H^3\iota/t}$.
    $Q_h(x_h, a_h) \leftarrow (1-\alpha_t)Q_h(x_h, a_h) + \alpha_t[r_h(x_h, a_h) + V_{h+1}(x_{h+1}) + b_t]$.
    $V_h(x_h) \leftarrow \min\{H, \max_{a' \in \mathcal{A}} Q_h(x_h, a')\}$.

1. The use of UCB exploration in the model-free setting allows for better treatment of uncertainties for different states and actions.

   UCB exploration: $O(\sqrt{H^4 SAT\iota})$, $\iota = \log(SAT/\delta)$

2. Using learning rate$\alpha_t = \frac{H+1}{H+t}$, instead of $1/t$ to obtain regret that is not exponential in H.

Aggregated Q-learning with Upper Confidence Bounds

Algorithm 1: AQ-UCB

Input: $\mathcal{S}, \mathcal{A}, H, \{\phi_h\}_{h=1}^{H}, s_1, K$
Input: positive constants $\{\beta_n\}_{n=1,2,\ldots}$
Define constants $\alpha_t \leftarrow (H+1)/(H+t),\ t = 1, 2, \ldots$
Initialize $N_h(m) = 0, \hat{Q}_h(m) = H$ for all $h \in [H]$ and $m \in [M]$
Randomly draw the first trajectory $s_1^0, a_1^0, \ldots, s_H^0, a_H^0$, where $s_1^0 = s_1$
for episode $k = 1, \ldots, K$ **do**
  for stage $h = 1, \ldots, H$ **do**
    $m \leftarrow \phi_h(s_h^{k-1}, a_h^{k-1})$
    $N_h(m) \leftarrow N_h(m) + 1$
    $\hat{V}_{h+1} \leftarrow \max_{a \in \mathcal{A}} \hat{Q}_{h+1}(s_{h+1}^{k-1}, a)$
    $\tilde{Q}_h(m) \leftarrow (1 - \alpha_{N_h(m)}) \cdot \hat{Q}_h(m) + \alpha_{N_h(m)} \cdot \left[r_h^{k-1} + \hat{V}_{h+1} + \beta_{N_h(m)} \cdot \frac{1}{\sqrt{N_h(m)}} \right]$
    $\hat{Q}_h(m) \leftarrow \min \left\{ \tilde{Q}_h(m),\ H \right\}$
  end for
  $s_1^k \leftarrow s_1$
  for stage $h = 1, \ldots, H$ **do**
    Take action $a_h^k \leftarrow \mathrm{argmax}_{a \in \mathcal{A}} \hat{Q}_h\left(\phi_h(s_h^k, a)\right)$
    receive reward r_h^k and next state s_{h+1}^k
  end for
end for
Output: the greedy policy with respect to $\{\hat{Q}_h\}_{h \in [H]}$

Aggregated Q-learning with Upper Confidence Bounds

- only has to maintain the values of $\left\{\hat{Q}_h\right\}_{h\in[H]}$ and $\{N_h\}_{h\in[H]}$
- if the computation of $\phi_h(s,a)$ takes $\mathcal{O}(1)$ time
  time complexity of AQ-UCB is $\mathcal{O}(HMK + HAK)$
  space complexity is $\mathcal{O}(HM)$.

Theorem

Suppose $\{\phi_h\}_{h\in[H]}$ is an ϵ-error aggregation of the underlying MDP. We have that, for any $\delta > 0$, if we run K episodes of algorithm AQ-UCB with

$$\beta_i = 2H^{\frac{3}{2}}\sqrt{\log\frac{HK}{\delta}} + \epsilon\cdot\sqrt{i}, \quad i = 1, 2 \ldots$$

then with probability at least $1-\delta$,

$$\begin{aligned}\text{Regret}(K) \leq & 24\sqrt{H^5 MK \log\frac{3HK}{\delta}} \\ & + 12\sqrt{2H^3 K \log\frac{3}{\delta}} \\ & + 3H^2 M + 6\epsilon\cdot HK\end{aligned}$$

- $\epsilon = 0$, $\tilde{\mathcal{O}}\left(\sqrt{H^5 MK}\right) = \tilde{\mathcal{O}}\left(\sqrt{H^4 SAT}\right)$, if $M = SA$, $T = HK$
- $\epsilon > 0$, per period performance loss of the policy that AQ-UCB ultimately outputs is $\mathcal{O}(\epsilon)$, which matches the per period loss lower bound $\Omega(\epsilon)$ established in Van Roy(2006) (Performance loss bounds for approximate value iteration with state aggregation).

Proof Outline: Notations

- let $\{\phi_h\}_{h\in[H]}$ be an ϵ-error aggregation ($\epsilon \geq 0$).
- $\hat{Q}_h^k(m)$: the value function estimate $\hat{Q}_h$ of aggregate state m, at the end of episode k, with $\hat{Q}_h^0(m) = H$.
- $\tilde{Q}_h^k(m)$: the uncapped value function estimate $\tilde{Q}_h$ of aggregate state m, at the end of episode k.

$$\hat{Q}_h^k(m) = \min\left\{\tilde{Q}_h^k(m), H\right\}$$

- $N_h^k(m)$: the number of visits to aggregate state m at stage h in the first k trajectories (indexed from 0 to $k-1$).
- $\tau_h^j(m)$: the episode index of the j -th visit to aggregate state m, at stage h.

Proof Outline: Notations

Simplified notations $\hat{Q}_h^k(s,a), \tilde{Q}_h^k(s,a), N_h^k(s,a)$ and $\tau_h^j(s,a)$ that represent $\hat{Q}_h^k(\phi_h(s,a))$ $\tilde{Q}_h^k(\phi_h(s,a))$, $N_h^k(\phi_h(s,a))$ and $\tau_h^j(\phi_h(s,a))$, respectively.
Recall that

$$\beta_i = 2H^{\frac{3}{2}}\sqrt{\log\frac{HK}{\delta}} + \epsilon\cdot\sqrt{i}, \quad i = 1, 2\ldots$$

$$\alpha_t = \frac{H+1}{H+t}, \quad t = 1, 2, \ldots,$$

Adopt the notations

$$\alpha_t^0 = \prod_{j=1}^{t}(1-\alpha_j), \ \alpha_t^i = \alpha_i \prod_{j=i+1}^{t}(1-\alpha_j)$$

Since $\alpha_1 = 1$, $\alpha_t^0 = 0$ and $\sum_{i=0}^{t}\alpha_t^i = 1$ when $t > 0$.

Proof Outline: On policy error analysis

- The uncapped value functions estimates

$$\tilde{Q}_h^k(m) = \alpha^0_{N_h^k(m)} \hat{Q}_h^0(m) + \sum_{j=1}^{N_h^k(m)} \alpha^j_{N_h^k(m)} \left[r_h^{\tau_h'(m)} + \hat{V}_{h+1}^{\tau h'(m)} \left(s_{h+1}^{\tau_h^j(m)} \right) + \frac{\beta_j}{\sqrt{j}} \right]$$

- On-Policy error:

$$\begin{aligned}
\hat{V}_h^k\left(s_h^k\right) - V_h^*\left(s_h^k\right) &\leq \hat{Q}_h^k\left(s_h^k, a_h^k\right) - Q_h^*\left(s_h^k, a_h^k\right) \leq \tilde{Q}_h^k\left(s_h^k, a_h^k\right) - Q_h^*\left(s_h^k, a_h^k\right) \\
&\leq \alpha^0_{N_h^k\left(s_h^k, a_h^k\right)} \cdot \left(H - Q_h^*\left(s_h^k, a_h^k\right)\right) \\
&+ \sum_{j=1}^{N_h^k\left(s_h^k, a_h^k\right)} \alpha_{N_h^k\left(s_h^k, a_h^k\right)} \left[r_h^{\tau_h^j\left(s_h^k, a_h^k\right)} + \hat{V}_{h+1}^{\tau_h^j\left(s_h^k, a_h^k\right)} \left(s_{h+1}^{\tau_h^j\left(s_h^k, a_h^k\right)} \right) \right. \\
&\left. + \frac{\beta}{\sqrt{j}} - Q_h^*\left(s_h^k, a_h^k\right) \right]
\end{aligned}$$

Proof Outline: On policy error analysis

$$
\begin{aligned}
=&\alpha^0_{N_h^k\left(s_h^k,a_h^k\right)} \cdot \left(H - Q_h^*\left(s_h^k, a_h^k\right)\right) \\
&+ \sum_{j=1}^{k_h^k\left(s_h^k,a_h^k\right)} \alpha^j_{N_h^k\left(s_h^k,a_h^k\right)} \left[r_h^{\tau_h^j\left(s_h^k,a_h^k\right)} + \hat{V}_{h+1}^{\tau_h^j\left(s_h^k,a_h^k\right)}\left(s_{h+1}^{\tau_h^j\left(s_h^k,a_h^k\right)}\right) + \frac{\beta}{\sqrt{j}} \right. \\
&\left. \quad - Q_h^*\left(s_h^{\tau_h^j\left(s_h^k,a_h^k\right)}, a_h^{\tau_h^j\left(s_h^k,a_h^k\right)}\right)\right] \\
&+ \sum_{j=1}^{k_h^k\left(s_h^k,a_h^k\right)} \alpha^j_{N_h^k\left(s_h^k,a_h^k\right)} \underbrace{\left[Q_h^*\left(s_h^{\tau_h^j\left(s_h^k,a_h^k\right)}, a_h^{\tau_h^j\left(s_h^k,a_h^k\right)}\right) - Q_h^*\left(s_h^k, a_h^k\right)\right]}_{\leq \epsilon}
\end{aligned}
$$

Proof Outline: On policy error analysis

$$
\begin{aligned}
\hat{V}_h^k\left(s_h^k\right) - V_h^*\left(s_h^k\right) =& \alpha^0_{N_h^k\left(s_h^k, a_h^k\right)} \cdot \left(H - Q_h^*\left(s_h^k, a_h^k\right)\right) \\
&+ \underbrace{\sum_{j=1}^{N_h^k\left(s_h^k, a_h^k\right)} \alpha^j_{N_h^k\left(s_h^k, a_h^k\right)} \left[\hat{V}_{h+1}^{\tau_h^j\left(s_h^k, a_h^k\right)}\left(s_{h+1}^{\tau_h^j\left(s_h^k, a_h^k\right)}\right) - V_{h+1}^*\left(s_{h+1}^{\tau_h^j\left(s_h^k, a_h^k\right)}\right)\right]}_{q_1} \\
&+ \underbrace{\sum_{j=1}^{N_h^k\left(s_h^k, a_h^k\right)} \alpha^j_{N_h^k\left(s_h^k, a_h^k\right)} \left[V_{h+1}^*\left(s_{h+1}^{\tau_h^j\left(s_h^k, a_h^k\right)}\right) - \mathrm{P}_h V_{h+1}^*\left(s_h^{\tau_h^j\left(s_h^k, a_h^k\right)}, a_h^{\tau_h^j\left(s_h^k, a_h^k\right)}\right)\right]}_{q_2} \\
&+ \underbrace{\epsilon + \sum_{j=1}^{N_h^k\left(s_h^k, a_h^k\right)} \alpha^j_{N_h^k\left(s_h^k, a_h^k\right)} \frac{\beta_j}{\sqrt{j}}}_{q_3}
\end{aligned}
$$

Proof Outline: Optimism Event $\mathcal{E}_{\text{opt}}$

By Azuma-Hoeffding inequality, with probability at least $1-\delta$, for all $h \in [H]$ and $k \in [K]$,

$$\left| \sum_{j=1}^{N_h^k(s,a)} \alpha_{N_h^k(s,a)}^j \left[V_{h+1}^* \left(s_{h+1}^{\tau_h^j(s,a)} \right) - \mathrm{P}_h V_{h+1}^* \left(s_h^{\tau_h^j(s,a)}, a_h^{\tau_h^j(s,a)} \right) \right] \right|$$
$$\leq \frac{2H^{\frac{3}{2}}}{\sqrt{N_h^k(s,a)}} \cdot \sqrt{\log \frac{HK}{\delta}}$$

Proof Outline: Optimism Event $\mathcal{E}_{\text{opt}}$

By Azuma-Hoeffding inequality, with probability at least $1-\delta$, for all $h \in [H]$ and $k \in [K]$,

$$\left| \sum_{j=1}^{N_h^k(s,a)} \alpha_{N_h^k(s,a)}^j \left[V_{h+1}^* \left(s_{h+1}^{\tau_h^j(s,a)} \right) - \mathrm{P}_h V_{h+1}^* \left(s_h^{\tau_h^j(s,a)}, a_h^{\tau_h^j(s,a)} \right) \right] \right|$$

$$\leq \frac{2H^{\frac{3}{2}}}{\sqrt{N_h^k(s,a)}} \cdot \sqrt{\log \frac{HK}{\delta}}$$

$\Rightarrow$

$$q_2 \leq \frac{2H^{\frac{3}{2}}}{\sqrt{N_h^k\left(s_h^k, a_h^k\right)}} \cdot \sqrt{\log \frac{HK}{\delta}}$$

Proof Outline: On-Policy Error

$$
\begin{aligned}
q_3 &= \sum_{j=1}^{N_h^k\left(s_h^k,a_h^k\right)} \alpha^j_{N_h^k\left(s_h^k,a_h^k\right)} \left(\frac{\beta_j}{\sqrt{j}} + \epsilon\right) \\
&= 2\epsilon + 2H^{\frac{3}{2}}\sqrt{\log\frac{HK}{\delta}} \cdot \sum_{j=1}^{N_h^k\left(s_h^k,a_h^k\right)} \frac{\alpha^j_{N_h^k}\left(s_h^k, a_h^k\right)}{\sqrt{j}} \\
&\le 2\epsilon + \frac{4H^{\frac{3}{2}}}{\sqrt{N_h^k\left(s_h^k, a_h^k\right)}} \cdot \sqrt{\log\frac{HK}{\delta}}
\end{aligned}
$$

Notice that on-policy error inequality is recursive. Summing both sides over $k = 1, \ldots, K$, we have

$$
\begin{aligned}
\sum_{k=1}^{K} \chi_h^k \le \sum_{k=1}^{K} \hat{Q}_h^k\left(s_h^k, a_h^k\right) - Q_h^*\left(s_h^k, a_h^k\right) \\
\le \frac{6H^{\frac{3}{2}}K}{\sqrt{N_h^k\left(s_h^k, a_h^k\right)}} \cdot \sqrt{\log\frac{HK}{\delta}} + 2\epsilon K + \sum_{k=1}^{K} \sum_{j=1}^{N_h^k\left(s_h^k,a_h^k\right)} \alpha^j_{N_h^k\left(s_h^k,a_h^k\right)} \cdot \chi_{h+1}^{\tau_h^j\left(s_h^k,a_h^k\right)}
\end{aligned}
$$

Proof Outline: On-Policy Error

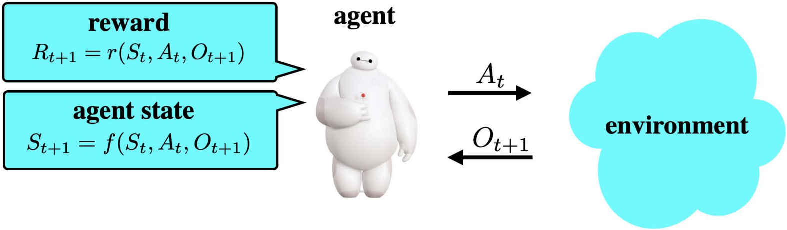
Notice that

$$\sum_{k=1}^{K} \sum_{j=1}^{N_h^k\left(s_h^k,a_h^k\right)} \alpha^j_{N_h^k\left(s_h^k,a_h^k\right)} \cdot \chi_{h+1}^{\tau_h^j\left(s_h^k,a_h^k\right)} \leq \sum_{k=1}^{K} \chi_{h+1}^k \cdot \sum_{t=N_h^k\left(s_h^k,a_h^k\right)+1}^{\infty} \alpha_t^{N_h^k\left(s_h^k,a_h^k\right)} \leq \left(1+\frac{1}{H}\right) \sum_{k=1}^{K} \chi_{h+1}^k$$

Then,

$$\begin{aligned}
\sum_{k=1}^{K} \chi_h^k &\leq \sum_{k=1}^{K} \hat{Q}_h^k\left(s_h^k, a_h^k\right) - Q_h^*\left(s_h^k, a_h^k\right) \\
&\leq \frac{6H^{\frac{3}{2}}K}{\sqrt{N_h^k\left(s_h^k, a_h^k\right)}} \cdot \sqrt{\log \frac{HK}{\delta}} + 2\epsilon K + \left(1+\frac{1}{H}\right) \sum_{k=1}^{K} \chi_{h+1}^k.
\end{aligned}$$

RL with Agent State



RL with Agent State: Environment $(\mathcal{A}, \mathcal{O}, \rho)$

- $\mathcal{A}$ is a finite set of actions
- $\mathcal{O}$ is a set of observations
- ρ is a conditional observation distribution $\rho(O_{t+1} \mid O_t, A_t)$
- The agent has access to the history

$$H_t = (A_0, O_1, A_1, O_2, \ldots, A_{t-1}, O_t)$$

RL with Agent State: Agent $(\mathcal{S}, f, r, S_0)$

- $\mathcal{S}$ is a finite set of agent states
- $f : \mathcal{S} \times \mathcal{A} \times \mathcal{O} \mapsto \mathcal{S}$ is an agent state update function

$$S_{t+1} = f\left(S_t, A_t, O_{t+1}\right)$$

- $r : \mathcal{S} \times \mathcal{A} \times \mathcal{O} \mapsto [0, 1]$ is a reward function (reflects the agent's preferences over histories)

$$R_{t+1} = r\left(S_t, A_t, O_{t+1}\right)$$

- $S_0 \in \mathcal{S}$ is an initial agent state.

RL with Agent State: Algorithm

Algorithm 1 Optimistic Q-learning.

Input: s_0, f, r
initialize restart timestamps $T_0 = 0, T_k = 20 \times 2^k$
`env.init()`
$t = 0, k = 0, s = s_0$
while true **do**
  if $t = T_k$ **then**
    $\gamma \leftarrow 1 - 1/T_{k+1}^{\frac{1}{5}}$
    $Q(s,a) \leftarrow 1/(1-\gamma), N(s,a) \leftarrow 0, \forall s, a$
    $\alpha_\ell \leftarrow \frac{2+(1-\gamma)}{2+\ell(1-\gamma)}, \ell = 1, 2, \ldots$
    $\beta \leftarrow 4\sqrt{\log T_{k+1}}/(1-\gamma)^{\frac{3}{2}}$
    $k \leftarrow k + 1$
  end if
  sample $a \sim \mathrm{unif}(\arg\max_{a' \in \mathcal{A}} Q(s, a'))$
  $n = N(s,a) \leftarrow N(s,a) + 1$
  $o \leftarrow \texttt{env.exec}(a)$
  $s' \leftarrow f(s, a, o)$
  $\tilde{Q} \leftarrow r(s,a,o) + \gamma \cdot \max_{a' \in \mathcal{A}} Q(s', a') + \frac{\beta}{\sqrt{n}}$
  $Q(s,a) \leftarrow (1-\alpha_n) \cdot Q(s,a) + \alpha_n \cdot \tilde{Q}$
  $s \leftarrow s', \quad t \leftarrow t + 1$
end while

- if computation of f takes $O(1)$ time, time complexity is $O(\mathcal{A}T)$, space complexity is $O(\mathcal{S}\mathcal{A})$.
- For $T \geq 1$,

$$\begin{aligned}\mathbb{E}[\text{Regret}(T)] \leq &\left(85\sqrt{\mathcal{S}\mathcal{A}\log(4T)} + 5\tau_{\tilde{\pi}_*}\right) T^{\frac{4}{5}} \\ &+ (81\mathcal{S}\mathcal{A} + 18\log(T))\, T^{\frac{1}{5}} \\ &+ 15\Delta T + 2\tau_{\tilde{\pi}_*}^{5}.\end{aligned}$$

Thank You!