Simple Agent, Complex Environment: Efficient Reinforcement Learning with Agent State

Shi Dong, Benjamin Van Roy, and Zhengyuan Zhou
Presenter: Jing Dong

University of Michigan, Ann Arbor

March 24, 2021

Motivation

Simulated environment

1. Data are generally available.

Real environments

1. The environment complexity is often much greater than what can be handled with available data.

Motivation

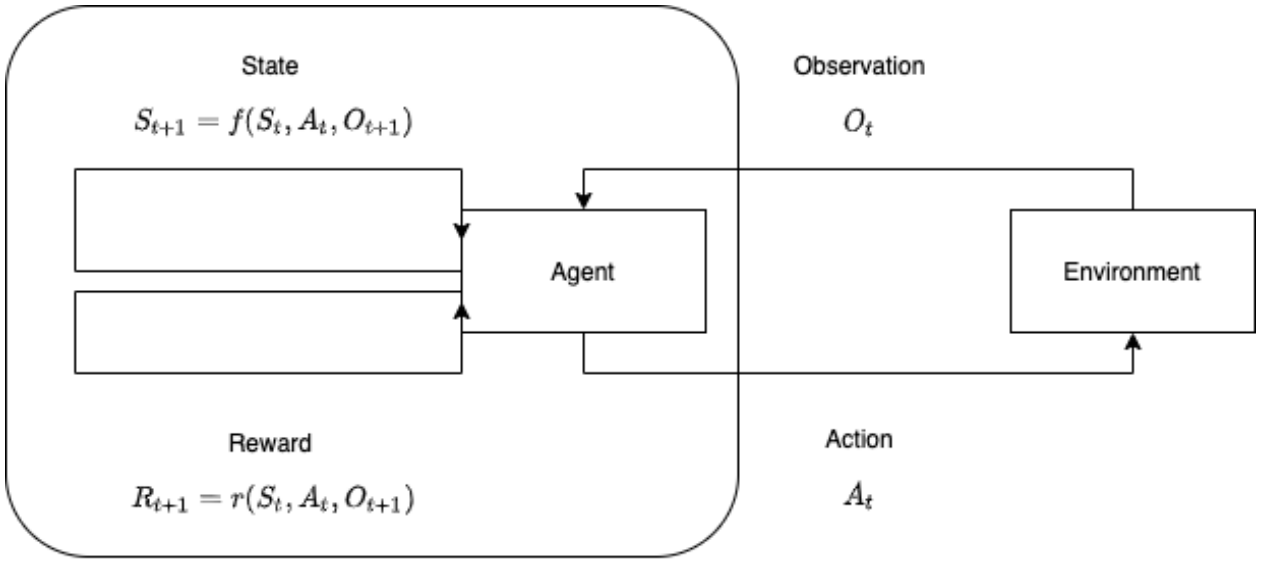
Can we design an agent to be efficient in a range of environment, where the environment may be more complex than the agent-environment interface complexity?

Agent-environment interface

Agent input

- Available actions $\mathcal{A}$,
- Possible observations $\mathcal{O}$,
- Agent states $\mathcal{S}$,
- Agent state update function $f : \mathcal{S} \times \mathcal{A} \times \mathcal{O} \longrightarrow \mathcal{S}$,
- Reward function $r : \mathcal{S} \times \mathcal{A} \times \mathcal{O} \longrightarrow [0, 1]$,
- initial agent state $s_0 \in \mathcal{S}$,

Agent-environment interface



History $H_t = \{A_i, O_i\}_{i \in t}$

Agent State and Policy

Consider two class of policy

- $\mathcal{P}$ based on current history
- $\tilde{\mathcal{P}}$ based on the current agent state

Clearly, $\tilde{\mathcal{P}} \subseteq \mathcal{P}$.

However,

1. History is unbounded,
2. policies in $\tilde{\mathcal{P}}$ require as input only the agent state, which can be updated incrementally and only demands a fixed amount of memory and per-timestep computation.

Performance measure and regret

Performance Measure

Performance of each policy π can be measured in terms of the expected average per-period reward

$$\lambda_\pi = \lim_{T \longrightarrow \infty} \inf \mathbb{E}_\pi [(1/T) \sum_{t=0}^{T+1} R_{t+1}] .$$

Regret

$$\text{Regret}(T) = \sum_{t=0}^{T-1} (\lambda_* - R_{t+1}) ,$$

where $\lambda_* = \sup_{\pi \in \mathcal{P}} \lambda_\pi$

Main result

$$\mathbb{E}[\text{Regret}(T)] = O((\sqrt{SA} + \tau_{\tilde{\pi}_*})T^{4/5} + SAT^{1/5} + \Delta T),$$

1. $\tau_{\tilde{\pi}_*}$ reward averaging time, which represents the time scale over which realized rewards should be averaged to accurately estimate $\lambda_{\tilde{\pi}_*}$,
2. Δ is a measure of achievable error in predicting optimal expected discounted return, with sufficiently large discount factors, based on agent state instead of history.

Main result

$$\mathbb{E}[\text{Regret}(T)] = O((\sqrt{SA} + r_{\tilde{\pi}_*})T^{4/5} + SAT^{1/5} + \Delta T)\,.$$

Agent is assured to operate reliably in arbitrarily complex environments. Notably, the regret bound does **NOT** depend on

1. the number of possible environment states,
2. environment dynamics,
3. reward averaging times of other policies,
4. mixing times of statistics beyond reward.

Main result

$$\mathbb{E}[\text{Regret}(T)] = O((\sqrt{SA} + r_{\tilde{\pi}_*})T^{4/5} + SAT^{1/5} + \Delta T)\,.$$

1. First result guaranteeing that an agent achieves near-optimal average reward in time polynomial in the number of agent states and actions and the reward averaging time of the optimal policy in the reference class;
2. First result demonstrating that a policy that achieves near-optimal average reward can be computed in time polynomial in the number of agent states and actions and an upper bound on reward averaging times.

Formal problem statement

Environment

The environment is epresented by tuple $(\mathcal{A}, \mathcal{O}, \rho)$, where $\mathcal{A}$ is a finite set of actions, $\mathcal{O}$ is a set of observations, and ρ is a conditional observation distribution.

After selecting action A_t, agent has access to the history $H_t = \{A_i, O_i\}_{i \in t}$. Let $\mathcal{H}_t$ be the set of all histories with duration t and $\mathcal{H} = \cup_{t=0}^{\infty} \mathcal{H}_t$ be the set of histories with finite duration, which is saying that the number of action and observations in each history is finite.

The transition to the next observation is governed by ρ. Specifically, the next observations takes value $o \in \mathcal{O}$ with probability $\rho(o|H_t, A_t)$. For all $o \in \mathcal{O}$ and any $h \in \mathcal{H}$, $s \in \mathcal{A}$, $\rho(o|h, a) \geq 0$ and $\sum_{o \in \mathcal{O}} \rho(o|h, a) = 1$.

Formal problem statement

Agent

The agent takes inputs as a tuple $(\mathcal{S}, f, r, S_0)$, where $\mathcal{S}$ is a finite set of agent states, $f : \mathcal{S} \times \mathcal{A} \times \mathcal{O} \longrightarrow \mathcal{S}$ is the agent state update function, $r : \mathcal{S} \times \mathcal{A} \times \mathcal{O} \longrightarrow [0, 1]$ is the reward function and $S_0 \in \mathcal{S}$ is the initial agent state.

The agent state evolves as

$$S_{t+1} = f(S_t, A_t, O_{t+1}) .$$

Formal problem statement

Agent

The agent state update function can also be redefined with a function that takes history as input

$$\psi_f(h) = f(\dots f(f(S_0, A_0, O_1), A_1, O_2) \dots, A_{t-1}, O_t)\,.$$

Policy

A policy π takes history $h \in \mathcal{H}$ and assigns a probability to each action $a \in \mathcal{A}$ such that $\sum_{a \in \mathcal{A}} = 1$.

Since an agent maintains agent state but not history in its memory, it can only execute policies for which action probabilities depend on history through agent state. We denote the set of such policies by

$$\tilde{\mathcal{P}} = \{\pi \in \mathcal{P} : \psi_f(h_1) = \psi_f(h_2) \longrightarrow \pi_{h_1} = \pi_{h_2}\}\,.$$

Value functions

The history value functions are defined as

$$V_\pi^\gamma(h) = \mathbb{E}_\pi[\sum_{t=0}^{\infty} \gamma^t R_{t+l+1} | H_l = h] \,,$$

where l is the duration of h. Similarly, we have

$$Q_\pi^\gamma(h, a) = \mathbb{E}_\pi[\sum_{t=0}^{\infty} \gamma^t R_{t+l+1} | H_l = h, A_l = a] \,.$$

Reward averaging time

Regardless of the initial history, as long as the agent follows policy π for τ_π/ϵ time steps, the expected average reward during this period lies in $[\lambda_\pi - \epsilon, \lambda_\pi + \epsilon]$. Formally, the reward averaging time is defined by

$$\tau_\pi = \sup_{h \in \mathcal{H}, T} \left| \mathbb{E}\left[\sum_{t=l}^{l+T-1} R_{t+1} | H_l = h \right] - \lambda_\pi \cdots T \right| .$$

Note

This implies that if $\lambda_\pi = \lambda_{\pi_*}$, then $\mathbb{E}[\text{Regre}(T, \pi)] \leq \tau_\pi$

Algorithm

Algorithm 2 Discounted Q-learning subroutine

Input: f, r, T, γ, β
`env.reset`(h)
$t = 0, s \leftarrow \phi_f(h)$
$Q(s,a) \leftarrow 1/(1-\gamma), N(s,a) \leftarrow 0, \forall s, a$
$\alpha_\ell \leftarrow \frac{2+(1-\gamma)}{2+\ell(1-\gamma)}$
while $t < T$ **do**
  sample $a \sim \text{unif}(\arg\max_{a' \in \mathcal{A}} Q(s,a'))$
  $n = N(s,a) \leftarrow N(s,a) + 1$
  $o \leftarrow$ `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

env.reset(h) resets the environment such that the initial history is h. Such a method is not always available in practice. The reason that we have it here is to demonstrate that our result holds regardless of the initial history, as long as the agent starts from the corresponding agent state.

Regret Bound

Theorem 2

If we run Algorithm 2 with

$$\beta = \frac{4\sqrt{\log(4T)}}{(1-\gamma)^{3/2}},$$

then for all $T \geq 1$ and initial history $h \in \mathcal{H}$, we have that

$$\mathbb{E}[\sum_{l=0}^{T-1} V_*(H_l) - V_{\pi_l}(H_l)]$$
$$\leq \frac{17}{(1-\gamma)^{3/2}} \cdot \sqrt{\mathcal{S}\mathcal{A}T\log(4T)} + \frac{13\Delta}{1-\gamma} \cdot T + \frac{\mathcal{S}\mathcal{A}}{(1-\gamma)^2},$$

where

$$\Delta = \sup\{|V_*(h_1) - V_*(h_2)| h_1, h_2 \in \mathcal{H}, \psi_f(h_1) = \psi_f(h_2)\}$$

Regret Bound: Notation

Let $V_t(s)$ be the value of agent state s at time t immediately before action A_t is taken and we use $V_t(h)$ as a shorthand for $V_t(\psi_f(h))$. Similarly, we use $Q_t(s,a)$ and $Q_t(h,a)$.
Further, for each $t \geq 0$, let π_t be the policy such that for each $h, a \in \mathcal{H} \times \mathcal{A}$,

$$(\pi_t)_h(a) = \mathbb{E}[\pi_h^{alg}(a)|H_t]\,.$$

Therefore we have

$$\mathbb{E}[V_{\pi^{alg}}(H_t)] = \mathbb{E}[\mathbb{E}[V_{\pi^{alg}}(H_t)|H_t]] = \mathbb{E}[V_{\pi_t}(H_t)]\,,$$

and

$$\mathbb{E}[V_{\pi_t}(H_t)] = \mathbb{E}[\mathbb{E}[\sum_{l=0}^{\infty}\gamma^l R_{t+l+1}|H_t]] = \sum_{l=0}^{\infty}\gamma^l \cdot \mathbb{E}[R_{t+l+1}]$$

Regret Bound: Some derivation

After some algebraic manipulations, we can have

$$\begin{aligned}\sum_{t=0}^{T-1} V_{\pi_t}(H_t) - V_{\pi^{alg}}(H_t) =& \frac{1}{1-\gamma}\Big(\sum_{t=0}^{T-1} V_t(H_t) - Q_*(H_t, A_t)\Big) \\ &+ \frac{\gamma}{1-\gamma}\big(V_T(H_T) - V_{\pi^{alg}}(H_{t+1})\big) \\ &- \frac{\gamma}{1-\gamma}\Big(\sum_{t=0}^{T-1} V_{t+1}(H_{t+1}) - V_*(H_{t+1})\Big) \\ &+ \frac{\gamma}{1-\gamma}\Big(\sum_{t=0}^{T-1}\big(PV_*(H_t, A_t) - V_*(H_{t+1})\big) \\ &+ \frac{\gamma}{1-\gamma}\Big(\sum_{t=0}^{T-1}\big(PV_{\pi^{alg}}(H_t, A_t) - V_{\pi^{alg}}(H_{t+1})\big)\,, \end{aligned} \tag{1}$$

where $PV(H_t, A_t) = \mathbb{E}_{H_{t+1}\sim P(H_t, A_t)} V(H_t)$.

Regret Bound: Probability Bound

Lemma 3. *If we run Algorithm 2 with*

$$\beta = \frac{4}{(1-\gamma)^{\frac{3}{2}}}\sqrt{\log\frac{2T}{\delta}},$$

then with probability at least $1-\delta$*, for all* $h \in \mathcal{H}$ *and* $t \geq 0$ *there is* $V_t(h) \geq V_*(h) - \Delta/(1-\gamma)$.

Regret Bound: Notation

Before we move on...
For agent state $s \in \mathcal{S}$ and action $a \in \mathcal{A}$, let $t_n(s,a)$ be the timestep corresponding to the n-th selection of action a in agent state s. We have that

$$\psi_f(H_{t_n(s,a)}) = s,\ A_{t_n(s,a)} = a\,.$$

Set $t_0(s,a)$ to be 0 for all (s,a) and let $n_t(s,a)$be the number of times that action a is selected in agent state s prior to, and not including, timestep t, i.e.

$$n_t(s,a) = \max\{n : t_n(s,a) < t\}\,.$$

Further, we define $\hat{Q}_n(h,a)$ to denote $Q_{t_n(\psi_f(h),a)}(h,a)$. Note that

$$\hat{Q}_n(H_t,A_t) = \max_{a\in\mathcal{A}} Q_{t_n(H_t,A_t)}(H_t,a) = V_{t_n(H_t,A_t)}(H_t)\,.$$

Regret Bound: Proof Sketch

Let event $\mathcal{G}$ be the event where lemma 3 holds, we have the following inequality conditioned on $\mathcal{G}$

$$
\begin{aligned}
V_t(H_t) - Q_*(H_t, A_t) \leq & \alpha_{n_t(H_t,A_t)}(Q_0(H_t, A_t) - Q_*(H_t, A_t))] \\
& + \Delta + \frac{3\beta}{\sqrt{n_t(H_t, A_t)}} \\
& + \gamma \cdot \sum_{i=1}^{n_t(H_t,A_t)} \alpha^i_{n_t(H_t,A_t)} \cdot (V_{t_i(H_t,A_t)}(H_{t_i(H_t,A_t)+1}) \\
& - V_*(H_{t_i(H_t,A_t)+1})) \,.
\end{aligned}
$$

Regret Bound: Proof Sketch

Because $Q_*(H_t, A_t) \leq V_*(H_t)$,

$$\begin{aligned} V_t(H_t) - V_*(H_t) \leq & \alpha^0_{n_t(H_t,A_t)} \cdot (\hat{Q}_0(H_t, A_t) - Q_*(H_t, A_t) \\ & + \Delta + \frac{3\beta}{\sqrt{n_t(H_t, A_t)}} \\ & + \gamma \cdot \sum_{i=1}^{n_t(H_t,A_t)} \alpha^i_{n_t(H_t,A_t)} \cdot (V_{t_i(H_t,A_t)}(H_{t_i(H_t,A_t)+1}) \\ & - V_*(H_{t_i(H_t,A_t)+1})) \,. \end{aligned}$$

Regret Bound: Proof Sketch

Similar to the Q learning proof, we can show that

$$V_t(H_t) \geq V_*(H_t) - \frac{\Delta}{1-\gamma}.$$

Therefore we have that

$$\begin{aligned}
&\sum_{t=0}^{T-1} \alpha^0_{n_t(H_t,A_t)} \cdot (\hat{Q}_0(H_t, A_t) - Q_*(H_t, A_t) \\
&\leq \frac{1}{1-\gamma} \cdot \sum_{t=0}^{T-1} \mathbb{I}(n_t(H_t, A_t) = 0) \\
&\leq \frac{\mathcal{SA}}{1-\gamma}.
\end{aligned}$$

Regret Bound: Proof Sketch

Let $X_t = V_t(H_t) - V_*(H_t) + \frac{\Delta}{1-\gamma}$, we have that

$$\begin{aligned}
\sum_{t=0}^{T-1} X_t &\leq \sum_{t=0}^{T-1} V_t(H_t) - Q_*(H_t, A_t) \\
&\leq \frac{\mathcal{S}\mathcal{A}}{1-\gamma} + 2\Delta T + \frac{3\beta}{\sqrt{n_t(H_t, A_t)}} \\
&\quad + \gamma \cdot \sum_{t=1}^{T-1} \sum_{i=1}^{n_t(H_t,A_t)} \alpha^i_{n_t(H_t,A_t)} \cdot X_{t_i(H_t,A_t)+1} \cdot \qquad (2)
\end{aligned}$$

The first inequality holds since $Q_*(H_t, A_t) \leq V_*(H_t)$

Regret Bound: Proof Sketch

Utilizing results from [CZS+], we have the following lemma.

Lemma 2. *The following properties hold for* α_k^i*:*

(a) For every $k \geq 1$, $\frac{1}{\sqrt{k}} \leq \sum_{i=1}^{k} \frac{\alpha_k^i}{\sqrt{i}} \leq \frac{2}{\sqrt{k}}$;

(b) For every $k \geq 1$, $\max_{i=1,\ldots,k} \alpha_k^i \leq \frac{4}{k(1-\gamma)}$ *and* $\sum_{i=1}^{k} (\alpha_k^i)^2 \leq \frac{4}{k(1-\gamma)}$;

(c) For every $i \geq 1$, $\sum_{k=i}^{\infty} \alpha_k^i = \frac{3-\gamma}{2}$.

Regret Bound: Proof Sketch

By using the property of the step sizes, we have that

$$\cdot \sum_{t=1}^{T-1} \sum_{i=1}^{n_t(H_t,A_t)} \alpha^i_{n_t(H_t,A_t)} \cdot X_{t_i(H_t,A_t)+1} \leq \frac{3-\gamma}{2} \sum_{t=0}^{T-1} X_t \,.$$

Combining with (2), we have

$$(1-\gamma \cdot \frac{3-\gamma}{2}) \sum_{t=0}^{T-1} X_t \leq \frac{\mathcal{S}\mathcal{A}}{1-\gamma} + 2\Delta T + \frac{3\beta}{\sqrt{n_t(H_t, A_t)}} \,.$$

Regret Bound: Proof Sketch

After some algebraic manipulations, we have

$$\sum_{t=0}^{T-1} \frac{1}{\sqrt{n_t(H_t, A_t)}} \leq 2\sqrt{\mathcal{S}\mathcal{A}T}\,.$$

Therefore

$$(1 - \gamma \cdot \frac{3-\gamma}{2}) \sum_{t=0}^{T-1} X_t \leq \frac{\mathcal{S}\mathcal{A}}{1-\gamma} + 2\Delta T + 6\beta\sqrt{\mathcal{S}\mathcal{A}T}\,.$$

Regret Bound: Proof Sketch

Because

$$1 - \gamma \cdot \frac{3-\gamma}{2} \geq \frac{1-\gamma}{2}.$$

we have

$$\sum_{t=0}^{T-1} X_t \leq \frac{2\mathcal{S}\mathcal{A}}{(1-\gamma)^2} + \frac{4\Delta T}{1-\gamma} + \frac{6\beta}{1-\gamma}.$$

Regret Bound: Proof Sketch

Thus,

$$V_t(H_t) - Q_*(H_t) \leq \frac{\mathcal{SA}}{(1-\gamma)} + 2\Delta \cdot T + 6\beta\sqrt{\mathcal{SA}T} + \frac{(3-\gamma)}{2} \sum_{t=0}^{T-1} (X_t - \frac{\Delta}{1-\gamma}) .$$

Regret Bound: Proof Sketch

By the decomposition that we had (1), we obtain

$$\begin{aligned}\sum_{t=0}^{T-1} V_t(H_t) - V_{\pi^{alg}}(H_t) \leq & 3/2 \sum_{t=0}^{T-1} X_t + \frac{1}{(1-\gamma)^2} + \frac{\mathcal{SA}}{1-\gamma} \\ & + 2\Delta T + 6\beta\sqrt{\mathcal{SA}T} \\ & + \frac{\gamma}{1-\gamma}\left(\sum_{t=0}^{T-1} PV_*(H_t, A_t) - V_*(H_{t+1})\right) \\ & + \frac{\gamma}{1-\gamma}\left(\sum_{t=0}^{T-1} PV_{\pi^{alg}}(H_t, A_t) - V_{\pi^{alg}}(H_{t+1})\right)\end{aligned}$$

Regret Bound: Proof Sketch

By using the Hoeffding inequality on the last two terms and conditioned on event $\mathcal{G}$, we have

$$\begin{aligned}
\mathbb{E}[\sum_{t=0}^{T-1} V_t(H_t) - V_{\pi_t}(H_t)] =& \mathbb{E}[\sum_{t=0}^{T-1} V_t(H_t) - V_{\pi^{alg}}(H_t)] \\
\leq& \mathbb{E}[\sum_{t=0}^{T-1} V_t(H_t) - V_{\pi^{alg}}(H_t)|\mathcal{G}] \\
\leq& \frac{17}{(1-\gamma)^{5/2}}\sqrt{\mathcal{S}\mathcal{A}T\frac{2T}{\delta}} \\
&+ \frac{13\Delta}{1-\gamma}T + \frac{6\mathcal{S}\mathcal{A}}{(1-\gamma)^2}
\end{aligned}$$

Substituting $\delta = 1/(2T)$ gives us the final answer.

Averaged return

Theorem 3. *Let $\pi' \in \mathcal{P}$ be an arbitrary policy. If we run Algorithm 2 with*

$$\beta = \frac{4\sqrt{\log(4T)}}{(1-\gamma)^{\frac{3}{2}}},$$

then for all $T > \log(T)/(1-\gamma)$,

$$\mathbb{E}\left[\sum_{t=0}^{T-1} \lambda_{\pi'} - R_{t+1}\right] \leq \frac{17}{(1-\gamma)^{\frac{3}{2}}} \cdot \sqrt{\mathcal{SA}T\log(4T)} + \left[13\Delta + (1-\gamma)\tau_{\pi'}\right] \cdot T + \frac{9\mathcal{SA} + 2\log(T)}{(1-\gamma)}. \tag{33}$$

Lemma 1. *For all $\pi \in \mathcal{P}$, $h \in \mathcal{H}$ and $\gamma \in [0,1)$,*

$$\left| V_\pi^\gamma(h) - \frac{\lambda_\pi}{1-\gamma} \right| \leq \tau_\pi.$$

Averaged return: Proof Sketch

$$
\begin{aligned}
\mathbb{E}\left[\sum_{t=0}^{T-1} V_{\pi'}^{\gamma}(H_t) - V_{\pi^{\text{alg}}}^{\gamma}(H_t)\right] &= \mathbb{E}\left[\sum_{t=0}^{T-1} V_{\pi'}^{\gamma}(H_t) - V_{\pi_t}^{\gamma}(H_t)\right] \\
&\geq \mathbb{E}\left[\sum_{t=0}^{T-1}\left(\frac{\lambda_{\pi'}}{1-\gamma} - \tau_{\pi'}\right)\right] - \mathbb{E}\left[\sum_{t=0}^{T-1}\sum_{\ell=0}^{\infty}\gamma^{\ell} R_{t+\ell+1}\right] \\
&= \mathbb{E}\left[\sum_{t=0}^{T-1}\sum_{\ell=0}^{\infty}\gamma^{\ell}\cdot\left(\lambda_{\pi'} - R_{t+\ell+1}\right)\right] - \tau_{\pi'}\cdot T \\
&= \mathbb{E}\left[\sum_{t=0}^{T-1}\frac{1-\gamma^{t+1}}{1-\gamma}\cdot\left(\lambda_{\pi'} - R_{t+1}\right)\right] \\
&\quad + \mathbb{E}\left[\sum_{t=T}^{\infty}\gamma^{t+1-T}\frac{1-\gamma^{T}}{1-\gamma}\cdot\left(\lambda_{\pi'} - R_{t+1}\right)\right] - \tau_{\pi'}\cdot T,
\end{aligned}
$$

The first equality is because of

$$\mathbb{E}[V_{\pi^{alg}}(H_t)] = \mathbb{E}[\mathbb{E}[V_{\pi^{alg}}(H_t)|H_t]] = \mathbb{E}[V_{\pi_t}(H_t)]\,,$$

and the second inequality is because of

$$\mathbb{E}[V_{\pi_t}(H_t)] = \mathbb{E}[\mathbb{E}[\sum_{l=0}^{\infty}\gamma^{l} R_{t+l+1}|H_t]] = \sum_{l=0}^{\infty}\gamma^{l}\cdot\mathbb{E}[R_{t+l+1}]\,,$$

and Lemma 1.

Averaged return: Proof Sketch

Since $|\lambda_\pi - R_t| \leq 1$,

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=T}^{\infty} \gamma^{t+1-T} \frac{1-\gamma^T}{1-\gamma} \cdot (\lambda_{\pi'} - R_{t+1})\right] &\geq -\sum_{t=T}^{\infty} \gamma^{t+1-T} \frac{1-\gamma^T}{1-\gamma} \\
&= -\frac{1-\gamma^T}{1-\gamma} \cdot \frac{\gamma}{1-\gamma} \\
&\geq -\frac{1}{(1-\gamma)^2}.
\end{aligned}$$

Let $T_0^\gamma = \lfloor \log(T)/(1-\gamma) \rfloor$, then for all $t > T_0^\gamma$,

$$\gamma^t \leq \gamma^{\frac{\log T}{1-\gamma}} \leq \left(\frac{1}{e}\right)^{\log T} = \frac{1}{T}.$$

Averaged return: Proof Sketch

This enables us to show that,

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=T_0^\gamma}^{T-1} \frac{1-\gamma^{t+1}}{1-\gamma} \cdot \left(\lambda_{\pi'} - R_{t+1}\right)\right] &= \mathbb{E}\left[\sum_{t=T_0^\gamma}^{T-1} \frac{1}{1-\gamma} \cdot \left(\lambda_{\pi'} - R_{t+1}\right)\right] - \mathbb{E}\left[\sum_{t=T_0^\gamma}^{T-1} \frac{\gamma^{t+1}}{1-\gamma} \cdot \left(\lambda_{\pi'} - R_{t+1}\right)\right] \\
&\geq \frac{1}{1-\gamma} \cdot \mathbb{E}\left[\sum_{t=T_0^\gamma}^{T-1} \left(\lambda_{\pi'} - R_{t+1}\right)\right] - \sum_{t=T_0^\gamma}^{T-1} \frac{\gamma^{t+1}}{1-\gamma} \\
&\geq \frac{1}{1-\gamma} \cdot \mathbb{E}\left[\sum_{t=T_0^\gamma}^{T-1} \left(\lambda_{\pi'} - R_{t+1}\right)\right] - \frac{1}{(1-\gamma)T} \cdot T \\
&\geq \frac{1}{1-\gamma} \cdot \mathbb{E}\left[\sum_{t=T_0^\gamma}^{T-1} \left(\lambda_{\pi'} - R_{t+1}\right)\right] - \frac{1}{(1-\gamma)}.
\end{aligned}$$

On the other hand,

$$\mathbb{E}\left[\sum_{t=0}^{T_0^\gamma - 1} \frac{1-\gamma^{t+1}}{1-\gamma} \cdot \left(\lambda_{\pi'} - R_{t+1}\right)\right] \geq -\frac{1}{1-\gamma} \cdot T_0^\gamma \geq -\frac{\log(T)}{(1-\gamma)^2}.$$

Averaged return: Proof Sketch

Leveraging the result from Theorem 2, that

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=0}^{T-1} V_{\pi'}^{\gamma}(H_t) - V_{\pi^{\mathsf{alg}}}^{\gamma}(H_t)\right] &\leq \mathbb{E}\left[\sum_{t=0}^{T-1} V_{*}^{\gamma}(H_t) - V_{\pi^{\mathsf{alg}}}^{\gamma}(H_t)\right] \\
&= \mathbb{E}\left[\sum_{t=0}^{T-1} V_{*}^{\gamma}(H_t) - V_{\pi_t}^{\gamma}(H_t)\right] \\
&\leq \frac{17}{(1-\gamma)^{\frac{5}{2}}} \cdot \sqrt{\mathcal{S}\mathcal{A}T\log(4T)} + \frac{13\Delta}{1-\gamma} \cdot T + \frac{6\mathcal{S}\mathcal{A}}{(1-\gamma)^2}.
\end{aligned}$$

Averaged return: Proof Sketch

As long as $T > T_0^{\gamma}$, there should be

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=0}^{T-1} \lambda_{\pi'} - R_{t+1}\right] &\leq \frac{17}{(1-\gamma)^{\frac{3}{2}}} \cdot \sqrt{\mathcal{SA}T\log(4T)} + 13\Delta \cdot T + \frac{6\mathcal{SA}}{(1-\gamma)} \\
&\quad + \frac{\log(T)+1}{(1-\gamma)} + 1 + (1-\gamma)\tau_{\pi'} \cdot T + T_0^{\gamma} \\
&\leq \frac{17}{(1-\gamma)^{\frac{3}{2}}} \cdot \sqrt{\mathcal{SA}T\log(4T)} + 13\Delta \cdot T + \frac{9\mathcal{SA}}{(1-\gamma)} \\
&\quad + \frac{2\log(T)}{(1-\gamma)} + (1-\gamma)\tau_{\pi'} \cdot T,
\end{aligned}$$

Optimistic Q learning

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 $a \sim \text{unif}(\arg\max_{a' \in \mathcal{A}} Q(s, a'))$
  $n = N(s,a) \leftarrow N(s,a) + 1$
  $o \leftarrow$ `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

Optimistic Q learning

Each epoch in Algorithm 1 is equivalent to the discounted Q-learning subroutine (Algorithm 2) with a different discount factor.
One more step in the proof is required, this is left to interested audience.