Algorithm 2 Infinite Q-learning with UCB-Hoeffding

1: **Initialized:** $Q(x,a) \leftarrow \frac{1}{1-\gamma}$ and $N(x,a) \leftarrow 0$ for all $(x,a) \in \mathcal{S} \times \mathcal{A}$.
2: **Define** $\iota(k) \leftarrow \log\left(SAT(k+1)(k+2)\right)$, $H \leftarrow \frac{\ln(2/(1-\gamma)\Delta_{\min})}{\ln(1/\gamma)}$, $\alpha_k = \frac{H+1}{H+k}$.
3: **for** step $t \in [T]$ **do**
4: Take action $a_t \leftarrow \operatorname{argmax}_{a'} Q(x_t, a')$, observe x_{t+1}.
5: $k = N(x_t, a_t) \leftarrow N(x_t, a_t) + 1$,
6: $b_k \leftarrow \frac{c_2}{1-\gamma}\sqrt{H\iota(k)/k}$, ▷ c_2 is a constant that can be set to $4\sqrt{2}$.
7: $\widehat{V}(x_{t+1}) \leftarrow \max_{a' \in \mathcal{A}} \widehat{Q}(x_{t+1}, a')$,
8: $Q(x_t, a_t) \leftarrow (1-\alpha_k)\, Q(x_t, a_t) + \alpha_k \left[r(x_t, a_t) + b_k + \gamma \widehat{V}(x_{t+1})\right]$,
9: $\widehat{Q}(x_t, a_t) \leftarrow \min\left\{\widehat{Q}(x_t, a_t), Q(x_t, a_t)\right\}$.

6 Algorithm for Discounted MDP

The pseudocode is listed in Algorithm 2. We acknowledge that Algorithm 2 relies on knowing a lower bound on $\Delta_{\min}$, and we leave it an open problem to develop a parameter-free algorithm.

7 Proofs for Discounted MDP

Notations Let $Q^t(s,a), \widehat{Q}^t(s,a), V^t(s), \widehat{V}^t(s), N^t(s,a)$ denote the value of $Q(s,a), \widehat{Q}(s,a), V(s), \widehat{V}(s), N(s,a)$ right before the t-th step, respectively. Let $\tau(s,a,i) := \max\{t : N^t(s,a) = i-1\}$ be the step t at which $(x^t, a^t) = (x,a)$ for the i-th time. We will abbreviate $N^t(x^t, a^t)$ for n^t when no confusion can arise. α_t^i is defined same as that in the finite-horizon episodic setting.

Proof of Theorem 3.2 We shall decompose the regret of each step as the expected sum of discounted gaps using the exact same argument as Eq (1), where the expect runs over all the possible infinite-length trajectories[5] taken by Algorithm 2:

$$(V^* - V^{\pi_t})(s_t) = \mathbb{E}\left[\sum_{h=0}^{\infty} \gamma^h \Delta(x_{t+h}, a_{t+h}) \middle| a_{t+h} = \pi_{t+h}\left(s_{t+h}\right)\right]. \tag{18}$$

Based on this expression, the expected total regret over first T steps can be rewritten as

$$\begin{aligned} \mathbb{E}[\text{Regret}(T)] &= \mathbb{E}\left[\sum_{t=1}^{T} \left(V^* - V^{\pi_t}\right)(x_t)\right] = \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{E}\left[\sum_{h=0}^{\infty} \gamma^h \Delta(x_{t+h}, a_{t+h})\right]\right] \\ &= \mathbb{E}\left[\sum_{t=1}^{T} \sum_{h'=t}^{\infty} \gamma^{h'-t} \Delta(x_{h'}, a_{h'})\right] \end{aligned} \tag{19}$$

Our next lemma is borrowed from Wang et al. (2019), which shows that Algorithm 2 satisfies optimism and bounded learning error with high probability. By abuse of notation, we still use $\mathcal{E}_{\text{conc}}$ to denote the successful concentration event in this setting. Recall that Algorithm 2 specifies $\iota(t) = \log\left(SAT(t+1)(t+2)\right)$ and $\beta_t = \frac{c_3}{1-\gamma}\sqrt{\frac{H\iota(t)}{t}}$.

Lemma 7.1 (Bounded Learning Error)**.** *Under Algorithm 2, event $\mathcal{E}_{\text{conc}}$ occurs w.p. at least $1 - \frac{1}{2T}$:*

$$\begin{aligned} \mathcal{E}_{\text{conc}} := \Bigg\{ &\forall (x,a,t) \in \mathcal{S} \times \mathcal{A} \times \mathbb{N}_+ : 0 \le \left(\widehat{Q}^t - Q^*\right)(x,a) \le \left(Q^t - Q^*\right)(x,a) \\ &\le \frac{\alpha_{n^t}^0}{1-\gamma} + \sum_{i=1}^{n^t} \gamma \alpha_{n^t}^i \left(\widehat{V}^{\tau(x,a,i)} - V^*\right)\left(x_{\tau(x,a,i)}\right) + \beta_{n^t} \Bigg\}. \end{aligned}$$

[5]For the convenience of analysis, when proving the upper bound we remove the constraint $t \in [T]$ in the for-loop of line 3. Instead, we allow the algorithm to take as many steps as we need, even yielding infinite-length trajectories.

Then we proceed to present an analog of Lemma 4.3 that bounds the weighted sum of learning error in the discounted setting.

Lemma 7.2 (Weighted Sum of Learning Errors)**.** *Under $\mathcal{E}_{\text{conc}}$, for every (C,w)-sequence $\{w_t\}_{t\geq 1}$, the following holds.*

$$\sum_{t\geq 1} w_t \left(\widehat{Q}^t - Q^*\right)(x_t, a_t) \quad \leq \quad \frac{\gamma^H C}{1-\gamma} + \mathcal{O}\left(\frac{\sqrt{wSAHC\iota(C)} + wSA}{(1-\gamma)^2}\right) \tag{20}$$

Proof. Recall that Lemma 7.1 bounds the learning error $\left(\widehat{Q}^t - Q^*\right)(x_t, a_t)$ on $\mathcal{E}_{\text{conc}}$. Thus we have:

$$\sum_{t\geq 1} w_t \frac{\alpha^0_{n^t}}{1-\gamma} \leq \sum_{t\geq 1} \mathbb{I}\left[n^t = 0\right] \cdot \frac{w}{1-\gamma} = \frac{SAw}{1-\gamma}; \tag{21}$$

$$\begin{aligned}
\sum_{t\geq 1} w_t \beta_{n^t} &= \sum_{s,a}\sum_{i} w_{\tau(s,a,i)}\beta_i = \frac{c_3\sqrt{H}}{1-\gamma}\sum_{s,a}\sum_{i} w_{\tau(s,a,i)}\sqrt{\frac{\iota(i)}{i}} \\
&\leq \frac{c_3\sqrt{H}}{1-\gamma}\sum_{s,a}\sum_{i=1}^{C_{s,a}/w} w\sqrt{\frac{\iota(C)}{i}} \leq \frac{2c_3\sqrt{H}}{1-\gamma}\sum_{s,a}\sqrt{C_{s,a}w\iota(C)} \qquad \left(C_{s,a} := \textstyle\sum_{i\geq 1} w_{\tau(s,a,i)}\right) \\
&\leq \frac{2c_3}{1-\gamma}\sqrt{SAHCw\iota(C)}; \qquad \text{(Cauchy-Schwartz inequality)}
\end{aligned} \tag{22}$$

Moreover,

$$\begin{aligned}
&\sum_{t\geq 1} w_t \sum_{i=1}^{n^t} \gamma \alpha^i_{n^t}\left(\widehat{V}^{\tau(x,a,i)} - V^*\right)\left(x_{\tau(x,a,i)}\right) \\
=&\gamma \sum_{t\geq 1}\left(\widehat{V}^t - V^*\right)(x_{t+1}) \sum_{i\geq n^t+1} w_{\tau(x_t,a_t,i)}\alpha_i^{n^t+1} \\
=&\gamma \sum_{t\geq 1}\left(\widehat{V}^{t+1} - V^*\right)(x_{t+1}) \sum_{i\geq n^t+1} w_{\tau(x_t,a_t,i)}\alpha_i^{n^t+1} + \gamma \sum_{t\geq 1}\sum_{i\geq n^t+1} w_{\tau(x_t,a_t,i)}\alpha_i^{n^t+1}\left(\widehat{V}^t - \widehat{V}^{t+1}\right)(x_t).
\end{aligned} \tag{23}$$

We let

$$\widetilde{w}_{t+1} := \sum_{i\geq n^t+1} w_{\tau(x_t,a_t,i)}\alpha_i^{n^t+1},$$

and further simplify (23) to be

$$\gamma \sum_{t\geq 2} \widetilde{w}_t \left(\widehat{V}^t - V^*\right)(x_t) + \gamma \sum_{t\geq 1} \widetilde{w}_{t+1}\left(\widehat{V}^t - \widehat{V}^{t+1}\right)(x_t). \tag{24}$$

For the first term of (24), we claim that $\{\widetilde{w}_t\}_{t\geq 2}$ is a $(C, (1+1/H)w)$-sequence. This can be verified by a similar argument to Ineq (15). We also have

$$\left(\widehat{V}^t - V^*\right)(x_t) = \widehat{V}^t(x_t) - V^*(x_t) = \widehat{Q}^t(x_t, a_t) - V^*(x_t) \leq \widehat{Q}^t(x_t, a_t) - Q^*(x_t, a_t) = \left(\widehat{Q}^t - Q^*\right)(x_t, a_t).$$

Therefore, the first term can be upper bounded by

$$\gamma \sum_{t\geq 2} \widetilde{w}_t \left(\widehat{V}^t - V^*\right)(x_t) \leq \gamma \sum_{t\geq 2} \widetilde{w}_t \left(\widehat{Q}^t - Q^*\right)(x_t, a_t). \tag{25}$$

For the second term of (24), we have the following observation:

$$\begin{aligned}\gamma \sum_{t\geq 1} \widetilde{w}_{t+1}\left(\widehat{V}^t - \widehat{V}^{t+1}\right)(x_t) &\leq \gamma(1+1/H)w\sum_s\sum_{t\geq 1}\left(\widehat{V}^t - \widehat{V}^{t+1}\right)(s) \\ &\leq \gamma(1+1/H)w\sum_s \widehat{V}^1(s) \leq \frac{\gamma(1+1/H)wS}{1-\gamma}. \end{aligned} \tag{26}$$

.

Plugging (25) and (26) back into (24), we obtain

$$\sum_{t\geq 1} w_t \sum_{i=1}^{n^t} \gamma\alpha_{n^t}^i \left(\widehat{V}^{\tau(x,a,i)} - V^*\right)\left(x_{\tau(x,a,i)}\right) \leq \gamma \sum_{t\geq 2}\widetilde{w}_t\left(\widehat{Q}^t - Q^*\right)(x_t,a_t) + \frac{\gamma(1+1/H)wS}{1-\gamma}. \tag{27}$$

Finally, combining (21), (22) and (27) and Lemma 7.1, we conclude that

$$\begin{aligned}&\sum_{t\geq 1} w_t\left(\widehat{Q}^t - Q^*\right)(x_t,a_t) \\ \leq& \sum_{t\geq 1} w_t\left(\frac{\alpha_{n^t}^0}{1-\gamma} + \beta_{n^t} + \gamma\sum_{i=1}^{n^t}\alpha_{n^t}^i\left(\widehat{V}^{\tau(s,a,i)} - V^*\right)\left(x_{\tau(s,a,i)+1}\right)\right) && \text{(Lemma 7.1)} \\ \leq& \frac{SAw}{1-\gamma} + \frac{2c_3}{1-\gamma}\sqrt{SAHCw\iota(C)} + \frac{\gamma(1+1/H)wS}{1-\gamma} + \gamma\sum_{t\geq 2}\widetilde{w}_t\left(\widehat{Q}^t - Q^*\right)(x_t,a_t) && (28)\end{aligned}$$

Note that the last term in Ineq (28) is another weighted sum of learning errors starting from step 2, where the weights form a $(C,(1+1/H)w)$-sequence. We can therefore repeat this unrolling argument for H times. Our choice of H in Algorithm 2 guarantees not only the bounded blow-up factor of weights, but also sufficiently small contribution of learning error after step H. In particular, we define a family of weights: when $h=0$, $\left\{w_t^{(h)}\right\}_{t\geq h+1} = \{w_t\}_{t\geq 1}$ is a (C,w) sequence; $\forall h\in[H]$ $\left\{w_t^{(h)}\right\}_{t\geq h+1}$ is a $\left(C,(1+1/H)^h w \leq ew\right)$ sequence. Note that our previous definition of $\widetilde{w}$ is exactly $w_t^{(1)}$.

$$\begin{aligned}&\sum_{t\geq 1} w_t\left(\widehat{Q}^t - Q^*\right)(x_t,a_t) \\ \leq\;& \sum_{h=0}^{H}\gamma^h\mathcal{O}\left(\frac{\sqrt{(1+1/H)^h wSAHC\iota(C)} + (1+1/H)^h wSA}{1-\gamma}\right) + \gamma^H\sum_{t\geq H+1} w_t^{(H)}\left(\widehat{Q}^t - Q^*\right)(x_t,a_t) \\ \leq\;& \mathcal{O}\left(\frac{\sqrt{wSAHC\iota(C)} + wSA}{(1-\gamma)^2}\right) + \frac{\gamma^H}{1-\gamma}\sum_{t\geq H+1} w_t^{(H)}. \end{aligned} \tag{29}$$

Using the fact that the weights after H unrolling $\left\{w_t^{(H)}\right\}_{t\geq H+1}$ is a $\left(C,(1+1/H)^H w \leq ew\right)$-sequence completes the proof. □

Note that we have clarified in Ineq (3) that on $\mathcal{E}_{\text{conc}}$ where optimism holds, sub-optimality gaps can be bounded by clipped learning error of Q-function. Again we divide its range $\left[\Delta_{\min}, \frac{1}{1-\gamma}\right]$ into disjoint subintervals and bound the sum inside each subinterval independently.

Lemma 7.3. *Let $N = \lceil \log_2(1/\Delta_{\min}(1-\gamma))\rceil$. On $\mathcal{E}_{\text{conc}}$, for every $n\in[N]$,*

$$\begin{aligned}C^{(n)} :=& \left|\left\{t\in\mathbb{N}_+ : \left(\widehat{Q}^t - Q^*\right)(x_t,a_t) \in \left[2^{n-1}\Delta_{\min}, 2^n\Delta_{\min}\right)\right\}\right| \\ &\leq \mathcal{O}\left(\frac{SA}{4^n\Delta_{\min}^2(1-\gamma)^5}\ln\left(\frac{SAT}{(1-\gamma)\Delta_{\min}}\right)\right).\end{aligned}$$

Again, based on Lemma 7.1, we prove Lemma 7.3 by choosing a particular sequence of weights.

Proof. For every $n \in [N]$, let

$$w_t^{(n)} \quad := \quad \mathbb{I}\left[\left(\widehat{Q}^t - Q^*\right)(x_t, a_t) \in \left[2^{n-1}\Delta_{\min}, 2^n\Delta_{\min}\right)\right], \tag{30}$$

then $C^{(n)} = \sum_{t=1}^{\infty} w_t^{(n)}$ and $\left\{w_t^{(n)}\right\}_{t\geq 1}$ is a $(C^{(n)},1)$-sequence. According to Lemma 7.2,

$$\begin{aligned}
\left(2^{n-1}\Delta_{\min}\right)\cdot C^{(n)} &\leq \sum_{t\geq 1} w_t^{(n)}\left(\widehat{Q}^t - Q^*\right)(x_t, a_t) \\
&\leq \frac{\gamma^H C^{(n)}}{1-\gamma} + \mathcal{O}\left(\frac{\sqrt{SAHC^{(n)}\iota(C^{(n)})} + SA}{(1-\gamma)^2}\right) \\
&= \frac{\Delta_{\min}}{2}C^{(n)} + \mathcal{O}\left(\frac{\sqrt{SAHC^{(n)}\iota(C^{(n)})} + SA}{(1-\gamma)^2}\right). \qquad \left(H = \frac{\ln\left(\frac{2}{\Delta_{\min}(1-\gamma)}\right)}{\ln\left(1/\gamma\right)}\right)
\end{aligned}$$

Now we proceed to solve the above inequality for $C^{(n)}$. For simplicity, let $\delta = 2^{n-2}\Delta_{\min}$ and $C^{(n)} = SAC'$. Then we have the following:

$$\begin{aligned}
\delta \cdot SAC' \quad &\leq \quad \left(2^{n-1} - \frac{1}{2}\right)\Delta_{\min}C^{(n)} \leq \mathcal{O}\left(SA\frac{\sqrt{HC'\iota(SAC')} + 1}{(1-\gamma)^2}\right), \\
\delta C' \quad &\overset{①}{\leq} \quad \mathcal{O}\left(\frac{\sqrt{C'}}{(1-\gamma)^{5/2}}\sqrt{\ln\left(SATC'\right)\ln\frac{1}{\Delta_{\min}\left(1-\gamma\right)}}\right), \\
C' \quad &\leq \quad \mathcal{O}\left(\frac{1}{\delta^2\left(1-\gamma\right)^5}\ln\frac{1}{\Delta_{\min}\left(1-\gamma\right)}\ln\left(SATC'\right)\right),
\end{aligned} \tag{31}$$

where ① comes from the definition $H = \frac{\ln(2/\Delta_{\min}(1-\gamma))}{\ln(1/\gamma)}$. Solving Ineq (31) yields

$$C' \leq \frac{1}{\delta^2\left(1-\gamma\right)^5}\ln\left(\frac{SAT}{\Delta_{\min}\left(1-\gamma\right)}\right).$$

Finally, substituting $C^{(n)} = SAC'$ and $\delta = 2^{n-2}\Delta_{\min}$ yields the desired formula. □

Proof of Theorem 3.2. We continue the calculation based on the regret decomposition in Eq (19). For every infinite-length trajectory traj $\in \mathcal{E}_{\text{conc}}$,

$$\begin{aligned}
\sum_{t=1}^{T}\sum_{h'=t}^{\infty}\gamma^{h'-t}\Delta(x_{h'}, a_{h'}|\text{traj}) &\overset{②}{=} \sum_{h=1}^{\infty}\Delta(x_h, a_h)\sum_{t=1}^{\min\{T,h\}}\gamma^t \leq \frac{1}{1-\gamma}\sum_{h=1}^{\infty}\Delta(x_h, a_h|\text{traj}) \\
&\overset{③}{\leq} \frac{1}{1-\gamma}\sum_{t\geq 1}\text{clip}\left[\left(\widehat{Q}^t - Q^*\right)(x_t, a_t|\text{traj})\middle|\Delta_{\min}\right] \\
&\overset{④}{\leq} \frac{1}{1-\gamma}\sum_{n=1}^{N}2^n\Delta_{\min}C^{(n)} \\
&\overset{⑤}{\leq} \mathcal{O}\left(\frac{SA}{\Delta_{\min}\left(1-\gamma\right)^6}\ln\left(\frac{SA}{p\epsilon\left(1-\gamma\right)\Delta_{\min}}\right)\right).
\end{aligned} \tag{32}$$

For the above inequalities, ② comes from an interchange of summations, ③ is by optimism of estimated Q-values, ④ is because we can add an outer summation over subintervals $n \in [N]$ and bound each of them by their maximum value times the number of steps inside. Finally, ⑤ follows directly from Lemma 7.3.

On the other hand, for trajectories outside of $\mathcal{E}_{\text{conc}}$, since sub-optimality gaps are upper bounded by $1/1-\gamma$, we have:

$$\sum_{t=1}^{T}\sum_{h'=t}^{\infty}\gamma^{h'-t}\Delta(x_{h'},a_{h'}|\text{traj}) \le \sum_{t=1}^{T}\sum_{h'=t}^{\infty}\gamma^{h'-t}\frac{1}{1-\gamma} \le \frac{T}{(1-\gamma)^2}. \tag{33}$$

Therefore, combining Ineq (32) and (33) gives us

$$\begin{aligned}\mathbb{E}[\text{Regret}(T)] &= \mathbb{E}\left[\sum_{t=1}^{T}\sum_{h'=t}^{\infty}\gamma^{h'-t}\Delta(x_{h'},a_{h'})\right] \\ &\le \mathbb{P}(\mathcal{E}_{\text{conc}})\cdot\mathcal{O}\left(\frac{SA}{\Delta_{\min}(1-\gamma)^6}\ln\left(\frac{SAT}{(1-\gamma)\Delta_{\min}}\right)\right) + \mathbb{P}\big(\overline{\mathcal{E}_{\text{conc}}}\big)\cdot\frac{T}{(1-\gamma)^2} \\ &\le \mathcal{O}\left(\frac{SA}{\Delta_{\min}(1-\gamma)^6}\ln\left(\frac{SAT}{(1-\gamma)\Delta_{\min}}\right)\right),\end{aligned} \tag{34}$$

where the last step is comes from $\mathbb{P}\big(\overline{\mathcal{E}_{\text{conc}}}\big)\le 1/2T$. Ineq (34) is precisely the assertion of Theorem 3.2. □

8 Difficulty in Applying Optimistic Surplus

The closest related work is by Simchowitz and Jamieson (2019) who proved the logarithmic regret bound for a model-based algorithm. Simchowitz and Jamieson (2019) introduced a novel property characterizing optimistic algorithms, which is called *optimistic surplus* and defined as

$$E_{k,h}(x,a) := Q_h^k(x,a) - \left[r_h(x,a) + P_h(x,a)^\top V_{h+1}^k\right]. \tag{35}$$

Under model-based algorithm with bonus term b_h^k, surplus can be decomposed as follows, where $\widehat{P}$ is the estimated transition probability:

$$E_{k,h}(x,a) = \left(\widehat{P}_h^\top(x,a) - P_h^\top(x,a)\right)V_{h+1}^* + \left(\widehat{P}_h^\top(x,a) - P_h^\top(x,a)\right)\left(V_{h+1}^k - V_{h+1}^*\right) + b_h^k.$$

The analysis of model-based algorithms is to first bound the regret $(V^* - V^{\pi_k})$ by a sum over surpluses that are clipped to zero whenever being smaller than some Δ-related quantities, then combine the concentration argument and properties of specially-designed bonus terms b_h^k to provide high probability bound for surpluses. However, for model-free algorithms, estimates of transition probabilities are no longer maintained, so $\widehat{P}_h$ is a one-hot vector reflecting only the current step's empirical sample drawn from the real next-state distribution. In this scenario, concentration argument of $\left(\widehat{P} - P\right)$ cannot give us $\log T$ regret.

Following the update rule of Q-learning with learning rate α_i and upper confidence bound b_i, the surplus becomes

$$E_{k,h}(x,a) = \alpha_t^0 H + \left(\sum_{i=1}^{t}\alpha_t^i V_{h+1}^{k_i}(x_{h+1}^{k_i}) - P_h(x,a)^\top V_{h+1}^k\right) + \sum_{i=1}^{t}\alpha_t^i b_i,$$

in which $t = n_h^k(x,a)$ is the number of times (x,a) has been visited, and $\alpha_t^i = \alpha_i\prod_{j=i+1}^{t}(1-\alpha_j)$ is the equivalent weight associated with the i-th visit of pair (x,a). This indicates that the surplus of an episode is closely correlated with estimates of value functions during previous episodes. The correlation makes the analysis more difficult. Therefore, we use a very different approach to analyze Q-learning in this paper.