We thank the reviewers for the insightful reviews and valuable suggestions. We address the comments as follows.

Reviewer #1:

Provide proof of Lemma 1: The proof of Lemma 1 uses induction. We will add the proof to the supplementary.

What is $R_{\max}$, a maximum value of the reward function R? Yes, as defined in Section 2.1.

Presentation issues of Algorithm 1: Many thanks! Yes, Algorithm 1 is meant for the synchronous case. We will clarify that lines 8 and 11 should be applied to all state action pairs. We will also specify the "converge" criterion in Algorithm 1, which can be, for example, when $\sum_{i=t-N}^{t} \left\|Q_i^A - Q_{i-1}^A\right\|_\infty < \varepsilon$ for a given $N \geq 1$ and $\varepsilon > 0$.

Discussion on the difficulties of analysis in the function approximation setting: This is a very important yet challenging problem. Even vanilla Q-learning with linear function approximation may not always converge and requires strong assumptions/conditions to converge, because the contraction property of the Bellman operator no longer holds in the function approximation setting. For double-Q algorithm, it is likely that neither of the two parameters (corresponding to the two Q-functions) converges, and even if they both converge, it is unclear whether they converge to the same point. Characterizing the conditions for these two interconnected stochastic processes to converge can be difficult.

Discussion on exploration policy in asynchronous version and Assumption 1: Yes, Assumption 1 is on a deterministic exploration policy for the simplicity of presentation of the key insights. The reason for $L \gg |S||A|$ is because in practice often some state-action pairs are repeatedly visited before the first visit of some other state-action pairs. $L = |S||A|$ can rarely happen in practice due to stochastic visits of state-action pairs. Assumption 1 can be replaced by a high probability requirement. Our analysis can accommodate such a relaxation by additionally dealing with a conditional probability event.

In Step I of Part I of proof of Theorem 1, provide intuition on the martingale difference sequence (MDS) z_t in the modeling of the interconnected stochastic processes: Indeed, interconnection of SAs introduces complication, as reflected by the non-MDS error sequence F_t (line 420 in supplementary). To handle the analysis of F_t, we decompose $F_t = F_t - \mathbb{E}(F_t|\mathcal{F}_t) + \mathbb{E}(F_t|\mathcal{F}_t) := z_t + h_t$, where we define $z_t := F_t - \mathbb{E}(F_t|\mathcal{F}_t)$ and $h_t := \mathbb{E}(F_t|\mathcal{F}_t)$. In this way, z_t is constructed to be an MDS by subtracting the conditional mean of F_t (a standard way to construct MDSs). The complication of non-MDS nature of F_t is captured by h_t, which we handle by exploiting a contraction-type property.

Minor comments: Many thanks for the suggestions. We will make the changes in the revision.

Reviewer #2:

Discussion on double Q-learning resolving overestimation: Our analysis of the interconnected SAs can also provide some insights into how double Q-learning resolves overestimation. High-level speaking, the convergence of $\|Q^A - Q^*\|$ is obtained with high probability given the condition that $\|Q^A - Q^B\|$ can converge. This suggests that neither Q^A nor Q^B can approach to Q^* alone too aggressively, which implies mitigation of overestimation.

Numerical results: We have obtained some numerical results and will include them in the revision.

Intuition on $\hat{\tau}_1$: We design the length of blocks in a recursive way, so that the ending time of the first block $\hat{\tau}_1$ determines the ending times of all following blocks. Then in our analysis, $\hat{\tau}_1$ is determined to guarantee that the distance between the two Q-functions can be bounded blockwisely with high probability.

Reviewer #3:

Lower bound on γ ($\gamma > 1/3$): In order for both $\{G_k\}$ and $\{D_k\}$ respectively serving as upper bounds on $\|Q^A - Q^B\|$ and $\|Q^A - Q^*\|$, we construct $G_k = \sigma D_k$ with $\sigma = \frac{1-\gamma}{2\gamma}$ in Proposition 2 (which may not be the only design to serve the purpose). Then, since the convergence of $\|Q^A - Q^*\|$ is conditioned on the convergence of $\|Q^A - Q^B\|$, we further need $D_k > G_k$, which requires $\gamma > 1/3$. This may not be an essential requirement, but is rather a technical assumption due to the techniques we use.

About linear learning rate: Great question! We are currently working on extending our techniques to the case with linear learning rate, and need to resolve some technical issues in order to obtain satisfactory results.

Reviewer #4: **Connection and difference with Even-Dar and Mansour (2003):** The connection lies in that our work also used the blockwise analysis as in Even-Dar and Mansour (2003). The difference is that the analysis of double Q-learning requires to handle two interconnected SAs, and the analysis of single SA in Even-Dar and Mansour (2003) cannot be directly applied. Our technical novelty lies in designing the coupling relationship between two blockwise upper bounds and dealing with conditional bounds.

Definition between “synchronous" vs. “asynchronous": Thanks for pointing this out. We will fix the statement of the synchronous algorithm in the revision.