We thank the reviewers for very helpful comments. This letter addresses the **major questions** raised by the reviewers.

Learning rates. To address the reviewers' comments on learning rates, we will add results with *easy-to-implement learning rates*, without compromising sample complexities. Specifically, for some constant $c > 0$ let

$$\eta_t = \min\Big\{1, c\exp\Big(\Big\lfloor \log \frac{\log t}{\widehat{\mu}_{\mathsf{min},t}(1-\gamma)\gamma^2 t}\Big\rfloor\Big)\Big\} \qquad \text{(an epoch-based choice)} \tag{1}$$

which can be viewed as a "piecewise approximation" of the rescaled linear stepsizes $\eta_t = \min\big\{1, \frac{c\log t}{\widehat{\mu}_{\mathsf{min},t}(1-\gamma)\gamma^2 t}\big\}$. Here, $\widehat{\mu}_{\mathsf{min},t}$ is the minimum entry of certain empirical state-action visitation probability vector.[1]. Clearly, this choice does *not* rely on the mixing time t_{mix}, minimum state-action occupancy probability μ_{min}, and target accuracy ε. Encouragingly, our current theory can be easily extended to cover this easier-to-implement learning rate choice:

Theorem 5 (ℓ_∞ sample complexity for achieving ε accuracy)**.** *Consider asynchronous Q-learning with learning rates* (1)*. There exists some universal constant $C > 0$ such that: for any $0 < \delta < 1$ and $0 < \varepsilon \le \frac{1}{1-\gamma}$, one has $\|Q_T - Q^\star\|_\infty \le \varepsilon$ with probability at least $1-\delta$, provided that the sample size (or number of iterations) T obeys*

$$T \ge C\max\Big\{\frac{1}{\mu_{\mathsf{min}}(1-\gamma)^5\varepsilon^2}, \frac{t_{\mathsf{mix}}}{\mu_{\mathsf{min}}(1-\gamma)}\Big\}\cdot \mathsf{polylog}\Big(|\mathcal{S}|, |\mathcal{A}|, \frac{1}{1-\gamma}, T, \frac{1}{\delta}, \frac{1}{\varepsilon}\Big). \tag{2}$$

Similarly, our theory for variance-reduced Q-learning can also be extended to a stepsize that does not depend on t_{mix}. More specifically, this requires two changes: (1) the epoch length needs to keep increasing (i.e. at the end of every epoch, run $t_{\mathsf{epoch}} \leftarrow 2t_{\mathsf{epoch}}$); (2) set $\eta_t = \frac{c\log t_{\mathsf{epoch}}}{\widehat{\mu}_{\mathsf{min},t}(1-\gamma)t_{\mathsf{epoch}}}$. This can be analyzed via a similar argument.

Proof of Theorem 5. We sketch the proof for the piecewise choice (1), which follows easily from our Theorem 1.

1) Set $T_0 = T/2$. Given that $\eta_t \le 1$, it is easily seen that $\|Q_{T_0} - Q^\star\|_\infty \lesssim T_0$. To simplify presentation, we assume here that T is the point where η_t undergoes a change (we can easily cover general cases via epoch-based analysis).

2) Choose $\widetilde{\varepsilon}$ s.t. $\frac{\mu_{\mathsf{min}}(1-\gamma)T/2}{\log(\frac{|\mathcal{S}||\mathcal{A}|T/2}{\delta})\log\frac{T}{2}} = \frac{C}{(1-\gamma)^4\widetilde{\varepsilon}^2}$, which obeys $\widetilde{\varepsilon} \le \varepsilon$ under Condition (2). Combining the piecewise choice (1) and Condition (2) implies: $\eta_t \equiv \frac{c'}{\log(\frac{|\mathcal{S}||\mathcal{A}|T}{\delta})}\min\big\{(1-\gamma)^4\widetilde{\varepsilon}^2, \frac{1}{t_{\mathsf{mix}}}\big\}$, $\forall t \in [T_0, T]$, where c' is some constant.

3) With the above learning rate condition in mind, invoking Theorem 1 of our paper with initialization Q_{T_0} ensures that $\|Q_T - Q^\star\|_\infty \le \widetilde{\varepsilon} \le \varepsilon$ with probability at least $1-\delta$, provided that the sample size condition (2) holds. □

Specific questions by Reviewer 1: 1. *"Implementable learning rates"*: See our response above on "learning rates". While the constant c in η_t can also be specified explicitly (by using specific constants in Bernstein inequality, etc), we caution that such a theoretical constant might be overly conservative in practice, given that our theory focuses on orderwise sample complexity bounds and does not strive to sharpen the constant. We will clarify this in the revision.

2. *"Main contributions"*: In comparison to the state-of-the-art [33] which unveiled tight scaling w.r.t. the important factors $\frac{1}{1-\gamma}$ and $\frac{1}{\varepsilon}$, our main focus is towards sharpening the dependency on the problem dimension $|\mathcal{S}||\mathcal{A}|$ through improving the dependency on μ_{min}. Specifically, we improve prior sample complexity bound by a factor of $1/\mu_{\mathsf{min}} \ge |\mathcal{S}||\mathcal{A}|$. Given that $|\mathcal{S}||\mathcal{A}|$ is often enormous in practice, our theory potentially leads to a notable improvement.

Specific questions by Reviewer 2: 1. *"Dependency of stepsizes on t_{mix}"*: See the response above on "learning rates".

2. *"Bounds in expectation"*. A bound on expectation can also be extracted by (1) using the boundedness nature of the Q-update and (2) choosing δ to be sufficiently small. We will add this in the revision.

Specific questions by Reviewer 3: "*Asynchronous Q-learning vs. A3C*": We'd like to clarify a possible source of confusion due to the different use of terminology in two different topics. The word "asynchronous" in Q-learning was often used in classical Q-learning literature (e.g. Tsitsiklis [41]) to indicate that: at every time only a single state-action pair in the Q-function is updated. This is in stark contrast to another line of recent literature on asynchronous optimization, which studies asynchronous updates of multiple CPU threads in parallel/multi-agent optimization (for instance, the A3C paper uses asynchronous SGD to simultaneously deploy/coordinate multiple CPU threads). Hence, the two settings are indeed quite different, although the Q-learning algorithm studied here might be used in any single thread to perform some component of an RL algorithm. We will clarify this in the revision to avoid confusion.

Specific questions by Reviewer 5: 1. *"Dependency of stepsizes on t_{mix}"*: See the response above on "learning rates".

2. *"Advantages of model-free methods:"* Thanks for raising concerns about our statements on model-based vs. model-free RL. We will rephrase these statements in the revision based on the reviewer's suggestion.

[1] Using concentration bounds in the supplement, we can ensure $\widehat{\mu}_{\mathsf{min},t} = (1+o(1))\mu_{\mathsf{min}}$ for all $t \gtrsim t_{\mathrm{mix}}/\mu_{\mathsf{min}}$ (up to log factor).