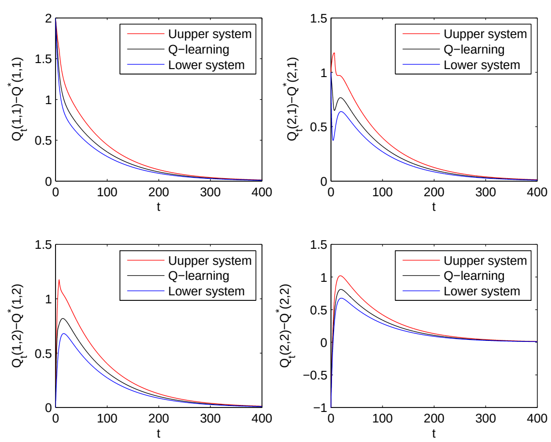
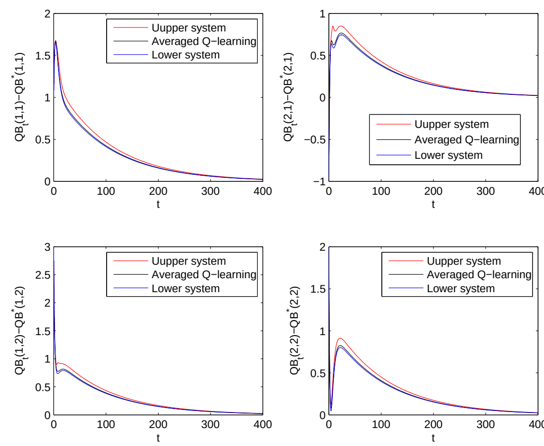
We thank all reviewers for their useful feedback and acknowledgement of our contribution. All comments will be addressed in greater details in the revision. We first answer some common questions brought up by reviewers.

Numerical illustrations: Thanks for the suggestion! We agree with reviewers that it is useful to provide some numerical evidence to illustrate the stability analysis. In Figures (a) and (b), we provide preliminary numerical illustrations of the associated ODE models (original affine switching systems, upper and lower comparison systems) of the *asynchronous Q-learning* and the *averaging Q-learning* on a toy MDP example with with $|S| = 2$ and $|A| = 2$. Our simulation empirically verifies the theory claimed in the paper. Richer numerical evidence will be included in the revision.



(a) Stability of asynchronous Q-learning



(b) Stability of averaging Q-learning

Assumptions: Given that this is the first work that bridges switching system theory with RL algorithms, we intentionally adopt simplified assumptions (such as i.i.d. assumption, orthogonal feature vectors) to avoid complications. Indeed, these assumptions are quite common in many seminal work in RL theory and can be relaxed. We will dedicate a discussion section on these assumptions and discuss potential relaxations or limitations.

Below we address the each reviewer's comments separately.

Response to Reviewer 1

Step-sizes: Using learning rates dependent on state-action observations may be useful in practice for small tabular MDPs; however, for modern developments in RL with function approximation, using learning rates independent of state-action observations is dominant in the literature.

Finite-sample guarantee: Our current framework only provides asymptotic convergence similar as most work on ODE analysis. Recent advances [Srikant & Ying, 2019; Hu & Syed, 2019; Chen et al., 2019; Wang & Giannakis, 2020; Devraj & Meyn , 2020] show promise in the derivation of non-asymptotic convergence rates using more sophisticated ODE analysis tools. We leave this extension for future investigation.

Response to Reviewer 2

Simulation/Assumptions: See discussions above.
Clarity of the paper: We will improve the presentation of the paper and avoid heavy notations.

Response to Reviewer 3

Numerical evidence. See discussions above.

Example. We used Example 1 as a *simple analytical illustration* of the tightness of these sufficient conditions. We agree that it might be too simple to justify the claim. We will consider nontrivial examples with interpretable feature matrices and verify these sufficient conditions numerically.

Response to Reviewer 4

Global Lipschitz continuity: This is not necessarily an assumption. We show that all the ODE models associated with Q-learning and its variants in this paper indeed satisfy the global Lipschitz continuity. Relaxing this condition into weaker ones would be promising to accommodate a wider array of RL algorithms, which we will pursue in the future.

Lemma 3: The spectrum is not well defined for switching systems because the subsystem matrices change. It is known that each subsystem matrix having negative spectrum does not guarantee the stability of the overall switching system. Lemma 3 is a particular *necessary and sufficient condition* for the stability of the overall switching system.