R1: No experimental evaluations: We did not focus on experiments as the goal was to demonstrate that our new technique for analyzing SA using the smoothed Lyapunov function is applicable for developing bounds for RL that can recover state-of-the-art bounds and enable new bounds for off-policy settings (please also see 'How optimal is the analysis').

R1: The results are only valid for pseudo-contraction updates: Our motivation for studying the SA algorithm in this setting is to analyze RL algorithms such as V-trace and Q-learning, which are known to have a contraction operator.

R1: Consider only martingale-difference noise, and finite state and action spaces: Since we consider the tabular method in RL, the underlying Markovian noise can be modeled by martingale differences. When using function approximation, Off-policy TD can potentially diverge [36]; studying it is one of our future direction.

R1: Title is too general: We will make the corresponding changes on the title.

R2: A number of assumptions made: Assumptions 2.1-2.3 in our paper are standard assumptions for studying SA algorithms involving a contraction operator, see Assumption 4.3 and Proposition 4.4 in [5]. Moreover, they are satisfied for many RL algorithms such as TD-learning and Q-learning, see Chapter 5 in [5]. Regarding the assumptions for the V-trace algorithm, they are from the original paper [17], and we do not make any additional assumptions.

R2: Novelty compared to prior work: Prior work studies either ℓ_2-norm contraction [5,10], or contraction w.r.t. $\| \cdot \|_\infty$ under the condition that the noise is uniformly bounded by a constant [3,4]. We establish convergence rate under general norm contraction and noise whose moments scale with the current iterate. (The lack of a smooth potential function for analyzing $\| \cdot \|_\infty$-contraction SA is a long-standing open problem, and is pointed out in [5], Sec 4.3 page 154). From a technical approach, we do not decompose the analysis into one for contraction and another for noise (as has been standard in prior works [3,4]). Our joint analysis of both is the key to our recursion (Proposition 2.1).

R2: How optimal is the analysis: The parameters in the Generalized Moreau Envelope can be tuned to tighten the bound. Though we do not have formal results on the optimality of our bounds, our approach based on a smooth Lyapunov function recovers existing state-of-the-art finite-sample bounds for Q-learning that show only a logarithmic dependence on the size of the state-action space [42] in a diminishing step-size regime, and improves over [3,4] in a constant step-size regime (see Appendix I of the supplementary materials for details).

R2: Simple ways to analyze SA with unbounded noise: When we have $\|w_k\| \leq c\|x_k\|$ (I think you are assuming that $x^* = 0$ is the fixed-point) for some small enough c, one can just use triangle inequality (even without taking expectation) to obtain a contractive recursion: $\|x_{k+1}\| \leq (1-\epsilon_k)\|x_k\| + \epsilon_k\|\mathbf{H}(x_k)\| + \epsilon_k\|w_k\| \leq (1-(1-\gamma-c)\epsilon_k)\|x_k\|$. However, in V-trace or Q-learning we have *affinely* increasing noise: $\|w_k\| \leq A(1+\|x_k\|)$, and as noted in Section 3.2 and Appendix I, the coefficient A is not small enough to apply this idea (and the affineness causes further difficulties).

R2: Analyze only a particular policy evaluation algorithm: Popular RL algorithms such TD(0), TD(n), TD(λ), Q-learning, and V-trace etc. can all be modeled by SA under contraction operator and martingale difference noise [5]. Thus our result is a broad tool to establish the finite-sample error bound of various RL algorithms.

R3: Synchronous V-trace: We agree with the reviewer that when performing asynchronous updates, there should be at least an additional factor of the dimension in the bound (indeed we see this in Q-learning). We will make this clearer in our paper. Studying convergence rates and concentration results for asynchronous V-trace is one of our future direction.

R3: Regarding $V_{\pi_{\bar{\rho}}}$ and V_{π}: When there is no clipping, we have $V_{\pi_{\bar{\rho}}} = V_{\pi}$. However, in this case the variance can be arbitrarily bad in the update, and is well recognized to be the key problem with off-policy methods. The goal of the V-trace algorithm is to reduce the variance by introducing the bias (i.e., introducing the clipper $\bar{\rho}$). By doing that, the variance is reduced to polynomial (quadratic) in $\bar{\rho}$. As for resulting bias (i.e., the gap between $V_{\pi_{\bar{\rho}}}$ and V_{π}), [17] discusses this at a high-level (Sec 4.1). A precise expression can be derived, but has complex dependencies on the behavior policy, target policy, and system parameters/dynamics. We will include this expression in the revised Supplementary material.

R3: Polynomial dependence on $\bar{\rho}$: We believe that a polynomial dependence on $\bar{\rho}$ is fundamental for any off-policy clipping based algorithms. Specifically, recall that if clipping is triggered, a sample is reweighted by a multiplicative factor of $\bar{\rho}$, which means that the signal and *noise* are both scaled by this factor. Further if this occurs for a constant fraction of time, the resulting noise variance scales order-wise as $\bar{\rho}^2$. Since we are looking at mean-square error, it is natural to expect a linear dependence on variance, which is what we see in our results. Thus, we believe that our results capture the correct scaling, and thus are significant for V-Trace.

R3: Minimizing the number of samples: For a given application, we can numerically optimize the parameters ($\bar{c}$, $\bar{\rho}$, T) to trade-off between contraction ratio and variance. We will discuss this in the revised draft.