We thank all reviewers for their careful reading of the paper, thoughtful feedback, and constructive suggestions. We look forward to revising the paper in light of your comments. Each reviewer's major comments are addressed below.

Reviewer 1. Thanks for your time and effort devoted to reviewing our submission, as well as for the positive comments on the analysis and exposition. Distinct novelties relative to Ref. [31] are: i) *Algorithm:* The present submission develops a novel decentralized multi-agent policy evaluation algorithm by wedding the merits of classic temporal-difference (TD) learning and contemporary gradient tracking, while Ref. [31] analyzed an *existing* distributed multi-agent TD learning algorithm; ii) *Analysis:* Our analysis is based on a single Lyapunov function unifying consensus and convergence while accounting for gradient tracking, yet [31] follows the standard path of proving error bounds for consensus and convergence separately; and, iii) *Bounds:* Our steady-state error bound does not depend on the number of agents N, matching that of centralized TD learning, while the steady error bound of [31] is N times *worse*. In a nutshell, although dealing with the same topic, our submission contributes a novel algorithm, a unifying analysis, and improved error bounds relative to [31]. Following your suggestion, [31] will be discussed more thoroughly in the revised paper.

Step-size. We will respectfully disagree that it "makes more sense to take a decaying step-size." In fact, *constant step-sizes* have been used and analyzed in the *seminal* contributions of TD learning [32, 1, 10, 38], as well as in recent ones [3, 5, 15, 31, 40, 41, 44]. Analyzing TD methods with constant step-sizes is critical, and serves as a first-step to analyze decaying step-sizes too. Yet, our algorithm and analysis can be generalized to accommodate decaying step-sizes. Due to space limitation, the focus of this paper was placed on analysis under *both IID and Markovian data*.

Bounds. In (28a)–(28c), $\lambda_1 < 0$ is the largest eigenvalue of the *negative* definite $\mathbf{A}$. If $|\lambda_1|$ is close to zero, it is clear from line 227 that the steady error bound constant c_1 is monotonically increasing in $|\lambda_1|$, so the closer $|\lambda_1|$ is to zero, the smaller the steady error bound is. On the other hand, the eigenvalues of $\mathbf{A}$ can be designed (to some extent) by properly selecting features $\{\boldsymbol{\phi}(s)\}_s$; moreover, the error bound can be made arbitrarily small by taking small enough step-size $\alpha > 0$. With these clarifications, it will be great if the reviewer can upgrade the evaluation of our work.

Reviewer 2. Comparison with paper #4026. We appreciate your time and effort put in this review, as well as the constructive feedback on our submission. As correctly pointed out, both papers deal with analysis of (decentralized) RL algorithms. The main novelties of this paper are the decentralized TD tracking technique, its corresponding unifying (consensus and convergence) analysis, as well as the resultant state-of-the-art error bound matching the centralized TD setup (improving upon existing distributed TD learning algorithms). Paper #4026 on the other hand, focuses on analysis of multi-agent control using distributed Q-learning when both constant and decaying step-sizes are employed (while gradient tracking is *not* used). Policy evaluation and control are two fundamental RL tasks, each of which presents unique challenges in algorithm design and analysis. We agree with your assessment that the two papers deal with related but different RL problems. Step-size. Please refer to lines 13-17 above.

Why plot error between DTDT and DTDL? This is because the optimum $\boldsymbol{\theta}^*$ is *unknown* in practice (in fact, finding $\boldsymbol{\theta}^*$ would require knowing the distribution of the underlying MDP, that is not available in the RL context). Fig. 1a shows that DTDT and DTDL find the same fixed point in the consensus space; while there is a big difference with respect to consensus and TD errors as depicted in Figs. 1b and 1c, corroborating the merits of TD tracking.

Improvement over [11, Thm. 2] by $1/N$. This is due to the fact that the constant $L := \sum_{v \in \mathcal{V}} L_v$ in β_2 and β_3 in [11, Thm. 2] is on the order of N (number of agents), as you can see from the sentence above Eq. (40) in the appendix of [11]; while the constants in Thm. 1 of the present submission do not depend on N.

Reviewer 3. We appreciate your time as well as positive appraisal of this submission. Extension to nonlinear case. There would be some difference regarding the theoretical claims between the linear and nonlinear value function approximation, such as the measure of the optimal solution, but the tracking technique as well as our convergence analysis can be generalized beyond the linear case.

Larger RL tests. As our bounds are (nearly) independent of problem size, the improvement of the proposed DTDT algorithm over DTDL still holds. Following your request, tests with large RL tasks will be added in the revised version.

Two $\mathbf{W}$-communications. Indeed, the improved error bound of DTDT comes at the price of doubling the communication requirements of DTDL. This communication overhead can be challenging in real-time and large-scale RL tasks. Developing communication-efficient alternatives is a fruitful future research direction.

Additional tests. Due to space limitations, only a single test was presented in the original submission. In the revised version, your insightful suggestions for including these additional numerical tests will be accommodated. Codes for the presented experiments are straightforward, so they were not included. However, per your request, codes for all experiments will be provided in the final paper for ease of reproducibility.

Reviewer 4. Thanks for the time and effort spent in reviewing this paper, as well as for recognizing its contributions. Thanks also for your careful reading, and pointing out the typos, which will be corrected in the revised version.