We thank all the reviewers for their careful feedback and will revise our paper accordingly. We start with a response addressing one common point raised by Reviewer 1 and Reviewer 3 regarding how to handle the case where $\sum_{i=1}^n p_i^\infty b_i \neq 0$. This case can be handled by a shifting argument if $\bar{A} := \sum_{i=1}^n p_i^\infty A_i$ is invertible. Notice the iteration $\xi^{k+1} = (I+\alpha A(z^k))\xi^k + \alpha b(z^k)$ can be rewritten as $\xi^{k+1} - \tilde{\xi} = \xi^k - \tilde{\xi} + \alpha\left(A(z^k)(\xi^k - \tilde{\xi}) + A(z^k)\tilde{\xi} + b(z^k)\right)$ for any $\tilde{\xi}$. Now we denote $\tilde{b}_i = A_i\tilde{\xi} + b_i$ and the above iteration just becomes $\xi^{k+1} - \tilde{\xi} = (I+\alpha A(z^k))(\xi^k - \tilde{\xi}) + \alpha\tilde{b}(z^k)$. When $\bar{A}$ is invertible, we can choose $\tilde{\xi} = -(\sum_{i=1}^n p_i^\infty A_i)^{-1}(\sum_{i=1}^n p_i^\infty b_i)$ such that $\sum_{i=1}^n p_i^\infty \tilde{b}_i = \sum_{i=1}^n p_i^\infty (A_i\tilde{\xi} + b_i) = 0$. Now we can directly apply the theory in our paper to obtain analytical formulas for $\mathbb{E}(\xi^k - \tilde{\xi})$ and $\mathbb{E}[(\xi^k - \tilde{\xi})(\xi^k - \tilde{\xi})^T]$, which eventually lead to formulas for $\mathbb{E}\xi^k$ and $\mathbb{E}[\xi^k(\xi^k)^T]$. A key question is when $\bar{A}$ will be invertible. Typically this can be guaranteed by some rank conditions on the feature matrix Φ whose i-th row is equal to $\phi(i)^T$. For TD(0), $\bar{A}$ is Hurwitz (and hence invertible) when the discount factor is smaller than 1, p_i^∞ is positive for all i, and Φ is full column rank. Such a fact is presented in the classic paper "An analysis of temporal-difference learning with function approximation" by Tsitsiklis and Van Roy. Similar facts can be found for other TD algorithms (e.g. see Assumption 2 and Appendix A in Ref [19] for DTD and ATD). The assumption that Φ is full column rank is standard and states that any redundant features have been removed. Reviewer 1 is correct in that a discount factor is needed. In our paper, the calculation of θ^* for TD(0) already involved such a shifting argument, and the condition $\sum_{i=1}^n p_i^\infty b_i = 0$ is enforced for Equation (13) due to the fact that the projected Bellman equation and the equation $\sum_{i=1}^n p_i^\infty b_i = 0$ are equivalent for TD(0). Notice θ^* only solves the projected Bellman equation and does not minimize the mean-square Bellman error. When $\bar{A}$ is singular, we can slightly modify the input terms in (14) and (20) and directly obtain analytical formulas for (q^k, Q^k). However, there is no convergence guarantee for this case. Now we address specific reviewer comments below.

Response to Reviewer 1: In the above response, we have already discussed the validity of the assumption $\sum_{i=1}^n p_i^\infty b_i = 0$ for TD algorithms and how to shift terms for the case where $\sum_{i=1}^n p_i^\infty b_i \neq 0$. Now we discuss how to ensure the assumption that $\bar{A}$ is Hurwitz. This is a standard assumption required even by the basic ODE approach which is used to prove asymptotic convergence. This assumption can be guaranteed by some rank conditions on the feature matrix Φ. For example, when Φ is full column rank, $\bar{A}$ is Hurwitz for Equation (13). A reference for this is the classic paper "An analysis of temporal-difference learning with function approximation" by Tsitsiklis and Van Roy. Similar conditions for other TD algorithms can be found in Refs [19, 25]. We emphasize that our approach does not require any extra assumptions compared with the existing approaches. Finally, the "-" sign in Line 213 is due to the Hurwtiz assumption.

Response to Reviewer 3: We thank the reviewer for the constructive suggestions on how to improve the readability of the paper. We will revise the paper accordingly. Regarding the assumption $\mathbb{E}p_i^\infty b_i = 0$, please see our response at the beginning of this rebuttal. We also want to mention that one way of extending our approach for the infinite sample space is by using operator theory. In this case, we will have some infinite dimensional variants of (5) and (6). Now the iterations on q^k and Q^k are described by infinite dimensional linear operators instead of finite dimensional linear operators (which are just matrices). A rigorous treatment of such extensions requires heavy mathematical notation due to the use of spectrum theory of linear operators. We will outline such ideas (without giving details) in our revised draft.

Response to Reviewer 4: We agree that the new insights on TD learning brought by our analysis should be made more transparent. We will focus more on TD learning and improve the clarity accordingly. We do think that the reviewer has misunderstood our paper regarding its interpretability, significance, and originality. We will revise to make the following clarifications. Regarding **interpretability**, our results are not more difficult to interpret than the mean square error bound in Ref [23]. The trace of the covariance matrix will immediately give us the mean square error. Consequently, by substituting the expressions of Q^k into the equation in Line 141 of our paper, we will directly get exact formulas and related bounds for the mean square error at any step k. Regarding **significance**, our exact formulas do bring new insights compared with existing sample bounds. Ref [3] requires an extra projection step to handle the Markov noise, so now we mainly compare our results with Ref [23]. Firstly, based on Statement 2 of Theorem 2 in our paper, the covariance matrix (or the mean square error) has an exact limit. In contrast, Ref [23] only shows that the final mean square error is bounded above. Secondly, a fundamental question is how tight the bounds in Ref [23] are. Does there exist an ergodic Markov chain such that the resultant final mean square error actually scales on the order $O(\alpha^m)$ for some constant $m > 1$? Our theory states that the answer to this question is no. Our exact formulas for the convergence rate and the final limits of (q^k, Q^k) can not only provide an upper bound for the mean square error, but also directly lead to lower bounds. This justifies the tightness of the upper bounds in Ref [23]. Thirdly, for large α region, our theory states that the mixing rate of the underlying Markov chain z^k poses a fundamental limitation for the convergence rate of TD learning. Statement 3 in Theorem 2 of our paper exactly characterizes this effect, and we provided further discussions in the last paragraph of our main paper. Such a fact is not explained by the theory in Ref [23] which focuses on small α region. Our theory sheds new light on how to choose large α at the early phase of TD learning. Regarding **originality**, our paper is the first that uses MJLS theory to analyze learning algorithms. Although Ref [15] presents a jump system formulation for stochastic optimization in supervised learning, the noise model there is IID and MJLS theory is not used there. Our paper is the first one that really bridges "Markov" jump linear system theory with learning.