We thank all reviewers for their detailed feedback. We will be sure to address all questions and incorporate all suggestions in the final version of the paper. Please see individual responses below.

Reviewer 1: **1. "I am not convinced by the main motivation of this paper for decoupled or independent learning...".** Motivation for the strongly independent setting comes from a game theoretic perspective. We view each agent as a strategic party who makes a policy update to improve their own utility each epoch, rather than as an algorithm with the goal of solving a particular optimization problem. Developing such independent learning protocols is understood as an interesting and difficult problem in MARL (see Sections 3.2 & 3.4 of [73], as well as the open problems in Section 6 of [73] about policy gradient methods). The analogous question in the special case of matrix games has attracted much attention in a line of work starting from [16]. Moreover, our two-timescale approach does not require coordination: if we think of there being two types agents who play against each other, one of whom takes larger step sizes than the other, then our result shows that they will converge to a Nash equilibrium. If the agents are to use vanilla gradient updates each epoch, the two-timescale approach is also necessary: even in the special case of matrix games, if the agents use equal step sizes, then their policies will not converge (see item 3 for Reviewer 3).
2. "The technical result of this paper is also a little bit weak comparing to several recent results with coupled algorithms...". We will cite these works and compare to their quantitative bounds in the final version, but we emphasize that it is not clear whether any of these results can extend to the independent setting we consider. We view the question of incorporating optimism (i.e., via exploration bonuses) into the independent learning setting as an interesting direction for future work. However, we believe that providing guarantees for the vanilla policy gradient method—even though they depend on the mismatch coefficient—is quite valuable, given that this technique is widely used in practice.
3. "What is the main difference in proof machinery between this paper and [74]?..." There is not a typo on line 212. In the different setting of linear-quadratic games, [74] assumes that for each update step the max-player plays a best response to the min-player's policy, which is updated using gradient descent. We believe that this strategy also works for the setting we consider, but our main result (Theorem 1) requires a significantly more involved proof: Rather than using an exact best response, we have to show that if the max-player only performs a single gradient ascent update at each iteration, its policy stays "sufficiently close" to the best response to the min player's policy.

Reviewer 2: **1. Regarding the relation between our work and [2] and [38]:** The analysis for the two-timescale update in [38] relies on the assumption that $f(x, y)$ is concave with respect to the max-player's iterate y. Since concavity does not hold in stochastic games, our proof requires some new ideas: whereas [38] explicitly considers the movement over time of the points $y^\star(x_t) := \arg\max_y f(x_t, y)$ and uses the quantity $f(x_t, y^\star(x_t)) - f(x_t, y_t)$ as a potential function to control the sub-optimality of the iterates y_t, we do do not consider the points $y^\star(x_t)$ and must instead use the gradients of the Moreau envelope of the function $y \mapsto f(x_t, y)$ to play the role of a potential function (Lemmas 9 & 13). Our proof requires the addition of a few other technical contributions not present in [38], including a PL-type inequality for the max function $x \mapsto \max_y f(x, y)$ (Lemma 12).
2. "I wish the authors have provided the orders..." – We provide the explicit bound in Theorem 1a in Section B.2.
3. On the works which apply MD with optimism to omit concentrability – Please see item 2 for Reviewer 1.
4. "I was expecting the author to have the complete study..." – While we believe that our results stand on their own, the extension to linear models and beyond is fairly straightforward, and we plan to pursue this in future work.

Reviewer 3: **1. On the practicality of our approach and asymmetry:** The problem of proving convergence guarantees for independent policy gradient algorithms is an important theoretical problem, and our results—while asymmetric in nature—constitute the first polynomial-time convergence guarantees. Even proving polynomial-time convergence is a challenging problem that requires new ideas beyond previous work (see item 1 for Reviewer 2). Though the runtime of our algorithm may preclude immediate practical implications, we believe that our result can serve as a baseline for future work developing other algorithms, such as extragradient-based ones as we discuss in Section 5.
2. "One issue is that the dependence in ϵ and in the action-space A and B do not seem very natural..." The likely sub-optimal dependence on A, B, ϵ is an artifact of the two-timescale updates, which require that $\eta_x \ll \eta_y$. The fact that η_x is so small slows down convergence: even in the nonconvex-concave case ([38]), in the stochastic setting the convergence rate is $O(\epsilon^{-8})$; our rate $O(\epsilon^{-12.5})$ is slower because of additional complexities due to the non-concavity of $y \mapsto f(x, y)$ and the ε-greedy exploration. Again, our aim is to provide the first polynomial time guarantees, and we hope our work will serve as a baseline for future results developing tighter dependence on problem parameters.
3. "Can you further explain why a two time-scale approach is necessary?..." As we discuss at the end of Section 3, even in the setting of zero-sum matrix games (which is the special case of stochastic games with one state and stopping probability $\zeta = 1$), it is known that if both players use GDA with equal step sizes, then the dynamics may cycle and the quantity in (6) can fail to converge to 0 as $N \to \infty$. We believe that one way around two-timescale updates is for the players to instead use the extragradient update, as discussed in Section 5.1; though equation (EG) allows for different step sizes η_x, η_y, typically one sets $\eta_x = \eta_y$ for extragradient. We believe that proving convergence rates for extragradient (with $\eta_x = \eta_y$) in stochastic games is a difficult yet promising direction for future work.
4. "Regarding the concentrabilty coefficient assumption..." – please see item 2 for Reviewer 1.