We thank the reviewers for their valuable comments and for pointing out the typos and clarity issues. We will revise accordingly. We first explain the novelty and significance of our work and then answer the questions of each reviewer.

(Novelty and Significance.) Our novel bilevel optimization formulation of actor-critic works for general RL problems, which motivates our algorithm which solves the policy evaluation subproblem at a faster timescale. We seem to first establish the finite-time convergence of actor-critic for LQR with ergodic cost. Even for general RL problems, existing convergence analyses are asymptotic and based on ODE approximations. Moreover, we develop a novel theoretical analysis framework that decouples optimization problems faced by the actor and critic respectively. More importantly, although we focus on LQR, our analysis can be readily extended to general RL problems with other policy optimization methods for the actor (e.g. PPO and TRPO) and other policy evaluation evaluation methods (e.g. TD(0)). Applying our analysis framework to general RL problems, we can show that actor-critic converges to a stationary point when using compatible value function, which will be added in the revision. Moreover, our analysis of GTD seems the first discrete-time convergence guarantee for policy evaluation under the ergodic setting, where the Bellman equation takes a different form from the discounted setting. We will clarify our contributions in the revision.

Reviewer #1

(Assumption B.1.) Projection is only for the technical reason, which is used to obtain the sublinear convergence rate for GTD updates. The convergence of GTD can be established without explicit projection thanks to to the convex-concave structure. But more careful analysis are needed to obtain similar convergence rate without projection.

(Gaussian policy.) Due to the existence of the score function $\nabla_K \log \pi_K(u \,|\, x)$ in the policy gradient theorem, actor-critic algorithm requires a stochastic policy. Here we adopt Gaussian policy for simplicity. In general, we could let the policy be $u = -Kx + \epsilon$, where ϵ is a independent noise with zero mean and a known density. In this case, actor-critic also finds the optimal policy due to 1) the family of value functions are compatible to the policy parametrization and 2) any saddle point of $J(K)$ is the optimal policy (geometry of LQR). We will add a detailed discussion in the revision.

(Minor comments.) **(i) Numerical experiments.** We will add numerical experiments to illustrate the convergence rate of our algorithm. **(ii) Line 23.** In the general case, we can only show policy gradient converges to a stationary point of the objective. However, there is no guarantee on the performance of this learned policy, which can be arbitrarily worse compared with the optimal policy. Thus, characterizing the optimality of policy gradient is an open problem. **(iii) Line 63.** By "asynchrony" we mean the fact the online actor-critic involves coupled updates of the actor and the critic. From the perspective of the critic, it aims to evaluate the value function of the critic, which is also changing. This fact caused great trouble in the analysis of actor-critic. **(iv) Line 68.** $\mathrm{svec}(X)$ maps a symmetric matrix $X \in \mathbb{R}^{d\times d}$ to a vector in $\mathbb{R}^{d(d+1)/2}$. The definition is given in §1 and we will add more details in the revision. **(v) Dual ascent.** In the dual ascent view of actor-critic, the dual variable μ is a probability distribution over $\mathcal{X} \times \mathcal{U}$, which can be written as $\mu = \rho_\pi \otimes \pi$ for some policy π, where ρ_π is the stationary distribution over $\mathcal{X}$ induced by π. Directly optimizing the dual variable requires a strong simulator that is able to sample arbitrary state-action pairs.

Reviewer #4

(Discussion.) The significance of this work should be considered in the reinforcement learning field. Indeed, both our algorithm and theoretical framework can be directly applied to general reinforcement learning problems. Our GTD analysis can be readily applied to general RL. For these problems, we can establish the convergence of actor-critic to the stationary point of $J(\pi_\omega)$. Also see **(Novelty and Significance.)**.

(i) Line 132. The papers cited in Line 132 have established that policy iteration, dynamic programming, and policy gradient are able to find the optimal policy of LQR. We will clarify this in the revision. **(ii) Line 152.** $\rho(A - BK)$ is the spectral radius of matrix $A - BK$, which is different from ρ_K, the stationary distribution of π_K. We will change the notation and clarify this. **(iii) Equation 3.1.** It should be I_k instead of I_d as $u \in \mathbb{R}^k$. Thanks for pointing out.

Reviewer #5

(Bilevel Perspective.) Here we use the bilevel optimization view to motivate our algorithm which updates the critic at the faster pace as it solves the lower-level subproblem. We do not contribute to general bilevel optimization but our focus is on RL: we provide a finite-time convergence analysis for actor-critic. Also see **(Novelty and Significance.)**.

(Eqn. (3.18).) In policy evaluation, at $(x, u) \in \mathcal{X} \times \mathcal{U}$, "Double sampling" refers to sampling two next states-action pairs (x_1, u_1) and (x_2, u_2) where x_1 and x_2 are independent and sampled from $P(\cdot \,|\, x, u)$. This is not practical. However, eqn.(3.18) only utilizes (x, u) and (x_1, u_1), which are two consecutive state-action pairs in a Markov chain.

(Off-Policy AC.) When the importance sampling ratio $\tau_K(x, u) = \pi_K(u \,|\, x)/\pi_b(u \,|\, x)$ satisfy certain regularity conditions, using the optimization formulation in (A.2), we can similarly establish the sublinear convergence rate of off-policy GTD, which yields an estimator of the natural policy gradient direction. Thus, we can similarly establish the convergence guarantees for off-policy AC. We will add these theoretical results in the revision.

(Comparison with Fazel et al.) Fazel et al. show that on-policy policy gradient converges to the optimal policy of undercounted LQR. We study actor-critic for LQR with ergodic costs and noisy state dynamics. They focus on deterministic policies and estimate the policy gradient via zeroth-order optimization. We adopt Gaussian policy with on- and off-policy actor-critic. Although our analysis of policy updates shares some similarity to Fazel et al., our analysis of critic GTD updates is novel and our theoretical framework can be extended to actor-critic for general RL problems.