Reviewer 1: Q1: I wonder if their analysis tricks of AC/NAC when applied to PG methods improve their guarantees too? If their analysis tricks do improve PG guarantees, how does it compare then? Is this a question of $1/B$ vs $1/\sqrt{B}$?

A1: Great question! This paper proposed two major tricks to improve the convergence rate of AC/NAC: **Trick I** of analysis of mini-batch sampling and **Trick II** of exploitation of self-reduced variance.

For **PG**, the variance error is not self-reduced, and hence trick II cannot improve its convergence rate. We next check that trick I does not improve its convergence rate either. Recall the best known convergence rate of PG is given in (Xiong et al. 2020) as $\mathcal{O}\big(\frac{1}{(1-\gamma)^2\sqrt{T}}\big)$. Thus, we require $T \geq \mathcal{O}\big(\frac{1}{(1-\gamma)^4\epsilon^2}\big)$ to achieve an ϵ-accurate stationary point. Note that PG algorithm further requires a Monte Carlo rollout with average length $L = \mathcal{O}\big(\frac{1}{1-\gamma}\big)$ to estimate Q-function for each sample. Thus, the sample complexity of PG is given by $TL = \mathcal{O}\big(\frac{1}{(1-\gamma)^5\epsilon^2}\big)$ as given in (Xiong et al. 2020). Now, applying trick I (minibatch sampling) to PG, we obtain the convergence rate of $\mathcal{O}\big(\frac{1}{(1-\gamma)^2 T}\big) + \mathcal{O}\big(\frac{1}{(1-\gamma)^2 B}\big)$. Thus, we require $T \geq \mathcal{O}\big(\frac{1}{(1-\gamma)^2\epsilon}\big)$ and $B \geq \mathcal{O}\big(\frac{1}{(1-\gamma)^2\epsilon}\big)$ to achieve an ϵ-accurate stationary point. Thus, the sample complexity of minibatch PG is $TBL = \mathcal{O}\big(\frac{1}{(1-\gamma)^5\epsilon^2}\big)$, which is the same as that of PG.

For **NPG**, it can also be checked that trick I does not improve its rate. Since NPG has self-reduced variance, trick II does improve the sample complexity $\mathcal{O}\big(\frac{1}{(1-\gamma)^8\epsilon^4}\big)$ of NPG given in (Agarwal et al. 2019) to $\mathcal{O}\big(\frac{1}{(1-\gamma)^7\epsilon^3}\big)$. This improved rate of NPG is still worse than the sample complexity $\mathcal{O}\big(\frac{1}{(1-\gamma)^4\epsilon^2}\big)$ of NAC given in our paper .

Reviewer 2: Q1: It would be interesting to complement the theoretical results with empirical results in toy problem.

A1: Thanks for the suggestion! We are working on experiments and will add these results to the revision.

Q2: For the error term that disappears with a larger mini-batch (line 211). Isn't this more of a variance error?

A2: Yes, this error term should be called as variance error. we will fix it in the revision.

Q3: Does Thm 1 depend on both assumptions or just assumption 2?

A3: Thm 1 is based on (a) Assumption 2 and (b) $\|\phi(s,a)\|_2 \leq 1$ for all (s,a) and $(\theta - \theta_\pi^*)^\top A_\pi(\theta - \theta_\pi^*) \leq -\lambda_A \|\theta - \theta_\pi^*\|_2^2$. Item (b) is stated in the paragraph before Thm 1, which has been justified in many previous studies.

Q4: In Assumption 1, should L_ϕ be L_ψ? **A4:** Yes, L_ϕ should be L_ψ.

Q5: What is $\mathbb{P}(s_t \in \cdot | s_0 = s)$ in Assumption 2? **A5:** $\mathbb{P}(s_t \in \cdot | s_0 = s)$ denotes the probability distribution of s_t conditioned on the initial state s_0. The notation is confusing and we will change it. Thanks for pointing it out!

Reviewer 3: Q1: The proof assumes a linear critic, which can introduce an approximation error in practice (see Theorem 2). It is unclear whether the gain in convergence speed outweighs the approximation error.

A1: Great point! Though linear critic can introduce an approximation error, a line of theoretical studies (including this work) naturally start from linear critic because it is analytically trackable. In fact, we find our analysis here can be extended to the nonlinear critic case (see our answer to Q2 below).

Q2: Can nonlinear SA be applied to a nonlinear critic as well?

A2: Great question! Yes. For a nonlinear critic, we can utilize the algorithm of nonlinear temporal difference learning with gradient correction (nonlinear TDC) to update critic's parameter, which can be analyzed by adapting our current analysis for nonlinear SA and existing technique for linear TDC. We can then incorporate the convergence analysis for the nonlinear TDC into our current analysis framework for AC/NAC to obtain the overall convergence analysis.

Q3: In practice, some have noticed improved convergence rates of NAC compared to AC (e.g. ACKTR v.s. A2C in [1]). However, this paper suggests a slower rate by a factor of $(1-\gamma)^{-2}$. (a) What could cause the difference and (b) how could the theory here guide development of deep RL algorithms?

A3: **(a)** Due to the different nature of AC and NAC, existing literature (including this paper) characterize their convergence rates by different metrics: AC by the gradient norm (as in Thm2), but NAC by the function value (as in Thm 3). Thus, the theoretical convergence rates of AC (Thm 2) and NAC (Thm 3) are not directly comparable. **(b)** Our theory here provide the following insights. First, our theory shows that NAC converges to a global optimal policy, while AC converges only to a first-order stationary point, which is likely a local optimum. This theoretical result explains practical observations that ACKTR achieves larger accumulated reward than AC, and in principle captures the advantage of NAC. Second, our theory also shows that mini-batch AC/NAC converges faster than single-sample AC/NAC, which suggests that practical implementation of AC/NAC can adopt minibatch and constant stepsize to achieve fast rate.