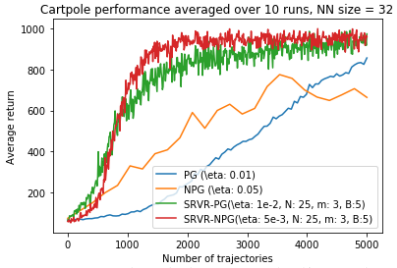
R1.1 *...these analysis mainly come from the existing work...the novelty is very limited.* We respectfully disagree. As pointed out by R2, R3, R4, and R5, this paper develops a novel global convergence *framework* that *unifies* the convergence of several policy gradient methods, whose novelty is summarized in Lines 210-212 and further explained in Lines 216-225. We proved the *global convergence* of SRVR-PG, for the first time; and improved the $\mathcal{O}(\varepsilon^{-4})$ sample complexity of NPG into $\mathcal{O}(\varepsilon^{-3})$. Our proposed SRVR-NPG has a better complexity than SRVR-PG (Remark 4.13).



R1.2 *...experimental results...* This paper focuses on laying the theoretical foundation for the global convergence of policy gradient methods, as [1,15,26,47]. Note that none of [1,15,26,47] has numerical results. We believed our theoretical contrition already has archival value. But still, we include a numerical result in the figure on the left. As this rebuttal will be archived in the NeurIPS proceeding, we assure this and more numerical results will be added, as well as more simulation details.

R1.3 *Reproducibility:* We believe that all of our theoretical claims have been proved. R1.4 *SRVR-NPG same complexity as SRVR-PG?* SRVR-NPG has a better iteration complexity, but it needs to sample trajectories to solve a subproblem at each iteration. Overall, its complexity has a better dependence on W and σ^2. R1.5 *Assumption 4.3 is a strict assumption...* As pointed out in Lines 199-201, this assumption is standard in the analysis of variance-reduced policy gradient methods, and it can be verified for Gaussian policies. Please refer to [34] for a detailed proof.

R1 As the main concerns are regarding the novelty and numerical validations, with our clarifications and new simulation results, we appreciate that the reviewer would re-evaluate our contribution, and change the scoring accordingly.

R2.1 *...other policy estimators...* Will add more discussions on this and open problems, when an extra page is available!

R3.1 *...It would be beneficial to include additional detail in the main paper...* Thanks for the suggestion. Proposition 4.5 applies the performance difference lemma and connects the global convergence rate with the stationary convergence rate. We will add more explanations in addition to Lines 210-212 and 240-243. R3.2 *...if these policies do not share the same support...* We agree that they should share the same support. Will add this. R3.3 *the global convergence...critically depends on ε_{approx}...* The richness of the function class explicitly occurs in the error bound, which will become very small or even zero under many common parametrizations, e.g., softmax, overparametrized neural nets, etc. This global convergence of RL cannot be provided by first-order guarantees, so it is much stronger than the latter in this sense.

R4.1 *...is the proved sample complexity tight?...the error due to the policy parametrization?* Interesting question. To the best of our knowledge, there isn't any lower bound for policy gradient methods under general policy parametrizations yet. We use $\varepsilon_{\text{approx}}$ to characterize the error due to policy parametrization (see Lines 204-208). R4.2 *...no experiment to demonstrate its empirical performance...* Please refer to R1.2 for some numerical results. R4.3 *...policy gradient methods exhibit high variance...inconsistent with the global convergence...?* Our results require sampling more trajectories per iteration than what's typically done in practice (e.g., $\mathcal{O}(\varepsilon^{-2})$ for PG). This will *stabilize* the performance, so we believe there is no inconsistency/counter-intuition from practice.

R5 Your detailed and thoughtful review is very helpful for us! Hope that our response will address your concerns.

R5.1 *...more discussion of the assumptions...how the results stated in earlier works can be translated...* We will make the presentation better, and make the translations more explicit. R5.2 *it would have been interesting to see some empirical work...whether the SRVR-NPG analysis is sub-optimal.* We have some numerical results in R1.2. Will add more to see if the analysis is tight or not. R5.3 *...more discussion...compare Assumption 2.1 with Assumption 6.2 in [1].* Will add more discussions. Specifically, the Assump 6.2 in the [1] (updated recently) implies our Assump 2.1. We found this independent but related finding quite interesting, and will discuss this. R5.4 *Section 2,3*: We have corrected the typos. R5.5 *...usually estimated via Monte-Carlo...* Sorry for the confusion. We will change the wording here. R5.6 *Assumption 4.1* In stochastic optimization, "variance" refers to the expectation of L^2 norm of the bias. Will clarify. R5.7 *Assumption 4.3* As mentioned in Lines 199-201, this assumption is standard in the analysis of variance-reduced policy gradient methods [34, 51, 52], and can be verified for Gaussian policies. Please refer to [34] for a detailed proof. R5.8 *Proposition 4.5* In Assumption 4.4, $\varepsilon_{\text{approx}}$ is an upper bound of *all* compatible function approximation error. R5.9 *Theorem 4.6* K should be $O((1-\gamma)^{-2}\varepsilon^{-2})$ and N should be $O(\sigma^2\varepsilon^{-2})$. Sorry for the confusion! Will define L_J. R5.10 *Remark 4.9* Yes, [1] does apply a small constant stepsize $\eta = \mathcal{O}(T^{-0.5})$, but with $T = \mathcal{O}(\varepsilon^{-2})$, not an absolute constant as ours. R5.11 *Lemma A.1* We believe that the calculation $\sum_{h=H}^{\infty} h\gamma^h = \left(H/(1-\gamma) + \gamma/(1-\gamma)^2\right)\gamma^H$ is correct. R5.12 *...an additional factor of 2...* There is a factor of 2 in the final line of (J.1) and (J.3), so in total we need 4. R5.13 *In (J.5),...choice of H...* We apologize for the confusion. The $\leq$ in the first equation of (J.5) should be $\geq$, it is a typo. We choose $H = \mathcal{O}\left(\log((1-\gamma)^{-1}\varepsilon^{-1})\right)$ so that the right-hand side is upper bounded by $\frac{1}{3}\left(\frac{\varepsilon}{3G}\right)^2$. R5.14 *...the new version of [1]...* Thanks for the notification. We will update our paper. Remarkably, with the new Assumption 6.5 of [1], we can simplify the analysis, and analyze the original NPG update without resorting to $\tilde{J}(\theta)$.