We would like to begin by thanking all the reviewers for their hard work in providing us with such insightful feedback. We are encouraged that the reviewers found our work to be simple and intuitive (**R1**, **R2**, **R3**), novel (**R1**, **R5**), and easy to implement (**R1**,**R2**). Several reviewers have also given credit to our work for sound theoretical claims/proofs (**R1**, **R5**) and thorough empirical evaluation (**R1**, **R3**). We will also be releasing the code for our implementation as it is our hope that our method not only provides a simple baseline for comparing new algorithms in constrained RL, but moreover makes it more accessible for researchers from other fields to apply RL to their own work.

Several reviewers noted that the guarantee in (9) may no longer hold post-approximation (**R1**, **R3**, **R5**). **R3** also pointed out that these approximations may prove to be ineffective in other applications. We acknowledge that these are valid concerns, but would also like to point out that the same can be said for most DRL methods. Our superior empirical performance compared to previous works show that the approximations we make are less destructive. As such, we do not believe our algorithm is less safe compared to CPO/PCPO (**R5**) which also makes extensive approximations. Furthermore, we are grateful to **R3** for noting that our approximations are both intuitive and effective.

We would like also to clarify the use of the indicator function in response to **R1** and **R2**. The indicator function enforces the constraint that π_θ is not too far from π_{θ_k}. This is also important because our method is a first-order method, so the approximations that we make is only accurate near the initial condition (i.e. $\pi_\theta = \pi_{\theta_k}$). We enforce this condition by ensuring $D_{KL}(\pi_\theta \parallel \pi_{\theta_k})$ do not diverge too much. The large distance between π_θ and π_{θ_k} doesn't mean that the distance between π_θ and π^* is large. At iteration k, before we make any update, $\pi_\theta = \pi_{\theta_k}$. As we make more gradient updates during iteration k, we expect π_θ to diverge from π_{θ_k} while becoming closer π^*. That is, the distance between π_θ and π_{θ_k} is *increasing*, but the distance between π_θ and π^* is *decreasing*.

Several reviewers recommended adding the constraint threshold values to the tables (**R1**, **R2**, **R3**). This will be done in our revision of the paper. Finally we would like to address other comments/concerns made by the reviewers.

R1 1) The goal of the MuJoCo environments is to train the agents to walk as fast as possible without falling over while not overexerting the joints. Hence the reward consists of multiple term which takes into account all such aspects. Our environment imposes a speed limit on the agents (which is reasonable in a safety-constrained setting) thus our policy forces the agent to optimize for the other terms in the reward (such as "stability" and torque applied to joints) while controlling for the speed. 2) In our subsequent revision we will explicitly write out the gradient terms for both PPO-L and TRPO-L. The reviewer is right in that the gradient for the cost term is very similar but the reward gradient term differ significantly since TRPO is a second-order method. 3) In our experiments, the random seeds determine both the initial weights of the neural nets and initial configuration of the environments.

R2 1) We have not had the opportunity to experiment on different constraint thresholds but we agree with the reviewer that these results would be interesting to see. We will be running these experiments and including the results in our subsequent revision. 2) In theory it is possible to extend FOCOPS to multiple constraints by introducing additional dual variables, we focused on the one constraint case since it results in cleaner maths and easier to interpret experiments. In our revision, we will make clearer the scope of our paper (single constraint). While FOCOPS like most similar work such as CPO focused on a single constraint, we concur that the multi-constraint case deserves further research.

R3 1) FOCOPS is an on-policy algorithm hence it inherits many of its flaws such as high sample complexity. We thank the reviewer for pointing this out. In our experience, learning constraint-satisfying policies from off-policy data is extremely challenging and deserves further research. 2) The reviewer is correct in pointing out that the theory of both FOCOPS and CPO assume the initial policy to be feasible. However in practice, the gradient update term increases the dual variable associated with the cost when the cost constraint is violated, this would result in a feasible policy after a certain number of iterations. We also observed that this is indeed the case with the swimmer environment.

R5 1) It is in general not computationally feasible to solve (1-3) directly therefore it would be difficult to compare its solution to FOCOPS. However it is possible to compare the gradient update term for CPO, which uses a second order approximation of (1-3) and the gradient update term for FOCOPS. We will add a brief discussion on this in our subsequent revision. 2) While we appreciate the novelty of PCPO's alternative two-step solution, empirically speaking PCPO does not seem to consistently beat CPO based on results reported in the original paper. To quote one of the meta-reviewers for PCPO from ICLR 2020: "The experimental evidence is a bit mixed, with the best of the proposed projections (based on the KL approach) sometimes beating CPO but also sometimes being beaten by it, both on the obtained reward and on constraint satisfaction". In contrast, FOCOPS outperformed CPO on all test environments. 3) In terms of computational speed, CPO takes one large gradient step while FOCOPS combines many smaller gradient steps using minibatches with early stopping. Due to the larger number of gradient steps, FOCOPS is in general slightly slower than CPO on most environments. However we found this difference to be marginal. 4) We would like to thank the reviewer for pointing us to the recent ICML 2020 paper from Stooke et al. and will add a brief discussion in our subsequent revision.