# **Better Exploration with Optimistic Actor-Critic**: Author Response

**All reviewers** Thanks for the feedback. As requested, we provide a plot measuring the sample efficiency gain (1) and additional ablations (Fig. 2 and 3). Also, OAC now supports deterministic policies as suggested by reviewer 1. While deterministic policies for exploration may appear surprising, deterministic OAC works because taking an action that maximises an upper bound of $Q^\pi$ is often a better choice than taking the action that maximises the mean estimate of $Q^\pi$. Results are in Fig. 4. Shaded bars denote one standard deviation (runs differ due to random initial state).

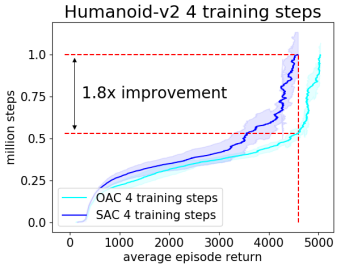

Fig. 1: Sample efficiency

**Reviewer 1** Thanks for the careful review. You are making several relevant points. You suggest extending OAC to support discrete, multi-modal and deterministic policies. We followed your third suggestion. We extended the scope of Proposition 1 slightly to include the Wasserstein distance, deriving an OAC variant that works with deterministic policies. We report the experimental results in Figure 4, where deterministic OAC beats deterministic SAC. You also suggested (points 1 and 2 in *detailed comments*) extensions to discrete and multi-modal policies. On discrete policies, the alternative to Proposition 1 would, as you say, no longer shift the policy mean but instead constrain policy change over the probability simplex. On multimodal-policies, extending OAC to Gaussian mixtures akin to Actor-Expert (Lim, 2018) would imply taking the KL divergence between mixtures. We agree that these extensions are interesting, but they would be hard to pack into a single submission. We will discuss them in the future work section and also relate our algorithm to Actor-Expert (Lim, 2018). In point 3, you suggested making an optimistic variant of TD3 or DDPG. We agree this would be informative, but we had a limited computation budget and chose SAC because of its performance on Humanoid. On your point 4, as you say, our results may not be groundbreaking but the difference is statistically significant and a step forward (Fig. 1). Also, thanks for flagging the TD3 results. There was a problem in our setup and the results are now 5K on Ant after 2.5M steps. Concerning the *small issues* part of the review, we will clarify description of TD3, fix the misnumbered equation and discuss the cost of computing the additional gradient (it is very small in practice).

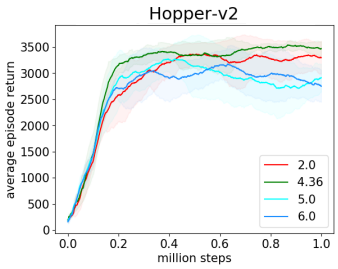

Fig. 2: Optimism ablation ($\beta_{\text{UB}}$)

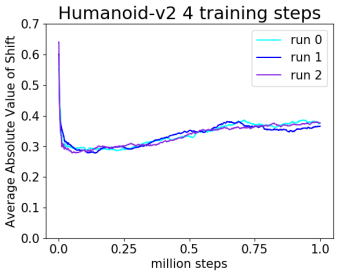

Fig. 3: Magnitude of shift.

**Reviewer 2** We appreciate the kind words. You are right when you say that OAC still needs many environmental interactions. However, using OAC vs SAC does make a meaningful difference. On Humanoid, OAC obtains a policy of same quality in 0.52M steps vs 1M for SAC (Figure 1). We agree that improving sample efficiency remains a challenge and hope that OAC paves the way for even better methods.

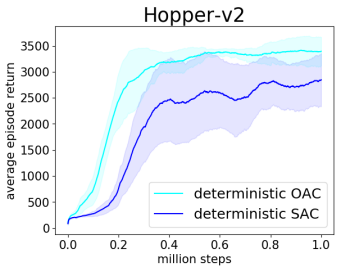

Fig. 4: Deterministic OAC

**Reviewer 3** Thanks for the feedback. In *detailed comments*, you ask why a spurious maximum of the lower bound leads to a policy with small covariance. Intuitively, this is because the actor finds a probability distribution that greedily maximises the critic lower bound. But a distribution that maximises a function is a point mass at the maximum of that function. Formally, as actor iteration progresses, the covariance $\Sigma$ can be modelled as $e^{Ht}$, where $H$ is the second order term in the Taylor expansion of the critic around the policy mean and $t$ is the iteration count. Near a maximum, $H$ is negative definite and we have $\Sigma \propto e^{Ht} \to 0$ as $t \to \infty$. We will include an extension of this argument in the paper. Your second point concerns how our Gaussian exploration policy avoids directional uninformedness. This is best seen in the figure on page 5 of the paper. While the exploration policy $\pi_E$ is symmetric around its own mean, it is not symmetric around the mean of the target policy $\pi_T$. We will make this clearer. Also, you requested a measurement of the directionality. We provide it in Fig. 3, which tracks the absolute magnitude in the difference between the mean of the exploration policy and target policy. We also preformed an ablation for optimism, shown in Fig. 2. The figure shows a sweet spot (the optimism value $\beta_{\text{UB}} = 4.36$ we used in the submission). About proposition 1, we will motivate $\Sigma_E = \Sigma_T$ more clearly. We will also expand the justification for this near line 450 of Appendix A. Also, you propose using $\max(\hat{Q}^1_{\text{LB}}, \hat{Q}^2_{\text{LB}})$ as the UCB. We in fact already do, i.e. $\max(\hat{Q}^1_{\text{LB}}, \hat{Q}^2_{\text{LB}}) = \mu_Q + \sigma_Q$, using notations from lines 154-158 and Appendix B. We will make this clearer. Also, as requested, we provide a plot measuring the number of steps to reach a given performance (Fig. 1). On your minor comments, we meant actor-critic in line 73. We will fix this as well as the typo, the misnumbered equation and the format of references.

## References

S. Lim, J. Ajin L. Le, Y. Pan, M. White *Actor-Expert: A Framework for using Action-Value Methods in Continuous Action Spaces*