We thank the reviewers for their interest in our work and their helpful comments. Please find our response below.

Comments relevant to all reviewers:

- In the three time scale procedure, faster time scales view slower time scales as static. This is why the fastest time scale is essentially solving a supervised learning problem over two static networks.

- The slowest time scale (delayed actor) is required both theoretically as well as empirically. Without it, convergence is not guaranteed and the algorithm becomes unstable.

Reviewer 1:

- You are correct in pointing out that an on-policy version of Algorithm 1 is not ensured to converge. DPO is an off-policy actor-critic framework which requires that all state action pairs are visited "enough" in order to ensure convergence, which is a theoretical assumption in various off-policy algorithms. We achieve this, similar to DQN, DDPG and TD3, by keeping an exploration strategy which does not decay to zero. We will emphasize this to avoid confusion in the paper.

- Your observation is correct, DPO and GAC are not perfectly aligned. DPO requires optimization in the distribution space, while GAC is a practical approximation of DPO in which optimization occurs in the space of parameters of the generative model. The DPO framework is a fundamental framework which can be extended in a similar way as Policy Gradient methods to bridge the gap between DPO and GAC.

- GANs and VAEs are definitely a valid choice for representing the policy, yet they have some pitfalls [1]. GANs pose the problem of learning a generative model and solving a two-player zero-sum game. This form of learning in itself is often unstable (resulting in mode-collapse) and still lacks theoretical guarantees and stability assurances. VAEs minimize the KL distance, as opposed to the p-Wasserstein distance, which has its own benefits. The quantile approach overcomes these issues by directly minimizing the p-Wasserstein distance using the quantile regression loss.

Reviewer 2:

- This is a good point. In practice, the three time-scale requirement is implemented using different learning rates so that the various elements converge at different rates. We follow a similar implementation method as in other actor-critic approaches, which are based on two timescales.

- *Derivation of policy distribution update:*
What we ultimately wish to have is an update similar to that of Policy Iteration (or more specifically, Approximate Policy Iteration), which conservatively updates the policy given a target policy π' as:

$$\pi_{k+1}(a|s) = (1 - \alpha_k)\pi_k(a|s) + \alpha_k \pi'(a|s).$$

Policy Iteration schemes use the target policy $\pi'(a|s) \in \arg\max_{a \in \mathcal{A}} r(s,a) + \gamma \sum_{s' \in \mathcal{S}} P(s'|s,a) v^{\pi_k}(s')$ in the exact case, or the ϵ-greedy target policy in the approximate case. Nevertheless, finding the $\arg\max$ is itself a hard problem in non-convex continuous regimes. Since finding the greedy action is complicated, it is reasonable to instead define the target policy (i.e., π') as a distribution over all improving actions. We denote this target policy by $\pi'(a|s) = \mathcal{D}^{\pi}_{I^{\pi}}(a|s)$. Finally, we take a gradient based approach using a distance metric over policies, d, yielding the DPO update rule

$$\pi_{k+1} = \Gamma\left(\pi_k - \alpha_k \nabla_\pi d(\mathcal{D}^{\pi_k}_{I^{\pi_k}}, \pi) \mid_{\pi = \pi_k}\right).$$

Reviewer 3: Thank you for pointing out some confusing explanations, we will make sure to clarify them in the paper.

- In Fig. 1a the intention is to compare optimization w.r.t. the policy itself (i.e., $\nabla_\pi v^\pi$) to optimization w.r.t. the parametrization $\nabla_\theta v^{\pi_\theta}$ (e.g., for Delta distributions θ represents the action, and for Gaussian distributions θ represents the mean μ). The former is what classical algorithms such as CPI (Kakade and Langford 2002) require, whereas the latter is what occurs in the standard policy gradient approaches. In Fig. 1a, the left (Π space) represents the ideal approach, whereas the right (Θ space) represents the sub-optimality which occurs when encountering a non-convex action-value function and optimizing with respect to the parametric distribution parameters (e.g., action).

- Regarding convexity, our intent was to show that the set Θ is not convex in a probabilistic sense. The set Π is the set of all probability distributions, whereas Θ is the span of probabilities distributions that π_θ can represent. Gaussian or Delta distributions are limited to their set, and thus can't ensure convergence to a global extrema. More specifically, $\alpha\delta_{\mu_1} + (1-\alpha)\delta_{\mu_2}$ means to play action $a_1 = \mu_1$ with probability α and $a_2 = \mu_2$ otherwise, but this distribution is not a Delta distribution, and therefore is not contained in Θ - the set of all Delta distribution functions. We will make this notation clear in the paper.

- Thank you for noting the mistake in lines 99 and 442 - we've updated the paper.

[1] Georg Ostrovski, Will Dabney, and Remi Munos. Autoregressive quantile networks for generativemodeling. *arXiv preprint arXiv:1806.05575*, 2018.