Response to Reviewer 1 Thank you for your constructive suggestions. We agree that Equation 7 with the divergence of state-action distribution (instead of only action distribution) is the core idea of the paper. The intuition for considering the discrepancy between state distributions, not only in action space, is to take future divergence into account. We agree that the algorithm should be better described as "ensures small and safe policy updates even with off-policy data". The paper will be updated according to your suggestions.

Response to Reviewer 2 We will revise the paper accordingly and improve the clarity of presentation.

Q: .. the Bregman divergence is argued to improve exploration. The intuition behind this is unclear to me.

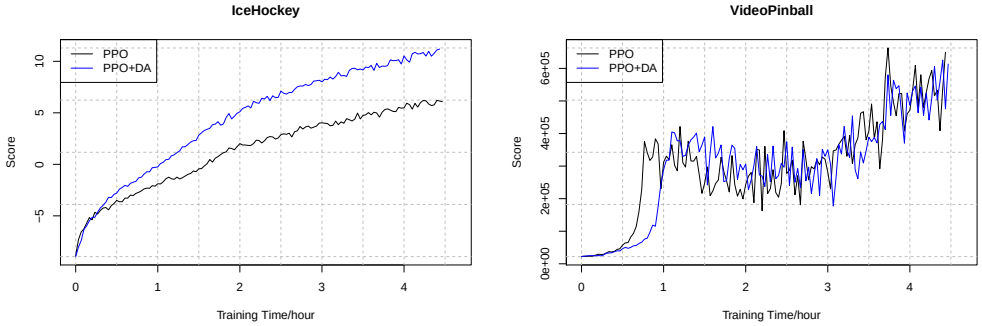
A: Intuitively, the divergence constraint can control the current policy π_t from going too far from previous policy π_{t-1} and retain the stochasticity of policy in the early stage. In addition, the divergence on state-action space also considers the discrepancy of policies on future states, thus encourages deeper exploration (line 41). We will investigate and emphasize the intuition of our method in the future version of paper.

Q: .. it is difficult to compare asymptotic performance with prior work, particular IMPALA.

A: The empirical difference between IMPALA and ours is mostly due to the number of actors used. As our focus is about the "data-scarce" scenarios, where the data generating speed is far more slower than the training speed (line 45), we use only 16 actors (line 432) to generate samples, while IMPALA use 210 (shallow) / 150 (deep) actors for DMLab and PBT for Atari (Sec.5.3.1). In general, higher score can be achieved by using more actors (more computational resource). Results with more actors and longer training time will be provided in the future version of paper.

Q: .. PPO baselines seem weaker and less stable than in the original PPO paper (e.g. ice hockey, video pinball).

A: The empirical performance may not be directly comparable with that in original PPO paper and IMPALA, due to the number of actors used, hyper-parameters, training infrastructure etc. We provide the empirical performance with 64 actors on IceHockey and VideoPinball below. The effect of number of actors will be discussed in the future version of paper.



Response to Reviewer 3 Thanks for the comments and pointing out typos. Source code will be released in the future.

O1: Our contribution is KL augmented return for policy optimization, corresponding to KL divergence on state-action space (line 35). The KL augmentation is between π_t and π_{t-1}, which is different from that in [Schulman et al. 2017a] (see line 191) and entropy in [Nachum et al. 2017]. The contribution will be pointed out more clearly and connection with related work will be discussed further in the future version of paper.

Q1: The effect of different factors can be measured in various aspects, e.g. by its stability as a proximal method (Sec 5.2.1), exploratory ability as suggested by divergence on future state distribution (Sec 5.2.2), etc. We will revise the questions more precisely and make the statements more clearly in Section 5.

Q2: a) The KL augmentation can be seen as a regularizer for proximal method, which can help escape local minimum. **b)** Intuitively, divergence measures the discrepancy from previous policy π_{t-1}, thus having infinitesimal effect during convergence (as $D(\pi_t||\pi_{t-1}) \to 0$ asymptotically); while entropy can be seen as measuring discrepancy from an uniform policy π_0 (i.e. $D(\pi_t||\pi_0)$, which may converge to a positive quantity, thus altering the learning objective).

C1: The mirror map formulation is related to the work of MPO and MARWIL and provided simply for completeness. We are also happy to remove this and just start from Equation 7.

C2 & C3: Thank you for pointing these out. We will update these parts accordingly in the next version of paper.

C4: We will elaborate on the meaning of x-y axis in our next version of paper.