We thank the reviewers for their insightful comments and suggestions. We hope that the rebuttal will clarify the issues.

Reviewer #1

1. (*Modified Algorithm of Definition 5.1*) The update rules of Definition 5.1 (leading to Algorithm 2 in the suppl. material) reduce to Definition 3.2 (leading to Algorithm 1) when $\mathcal{Q}_b = 0$ a.s. or $\beta_t = 0$. $\mathcal{Q}_b$ is used in the analysis as an additional source of uncertainty necessary to prove, together with a suitable choice of α_t and β_t, the PAC-MDP in the average loss for *any* tabular MDP. There are two reasons why Algorithm 1 should be preferred in practice. First, Algorithm 2 cannot be extended to continuous MDPs, as α_t and β_t are defined in terms of number of visits $n(s, a)$ (Equation (18) in the suppl. material), which can only be computed for finite MDPs. Second, as many provably efficient RL algorithms (e.g., MBIE or Delayed-QL), Algorithm 2 is extremely conservative, leading to very slow convergence. This is why most provably efficient RL algorithms, when used in practice, are run with non-theoretical values of hyperparameters (see Figures 2 and 3 of [2] for MBIE). Algorithm 1 can be seen as a "practical" version of Algorithm 2 in which α_t is treated as a normal hyper-parameter and $\beta_t = 0$. We will clarify this point in the final version.

2. (*Comparison with Distributional RL (DRL)*) For space reasons, we condensed the discussion in lines 119-121. While DRL models the distribution of the *return*, our WQL models the distribution of the Q-function estimate, which is defined as the *sample mean* of the *returns*. The two distributions are clearly related and both depend on the stochasticity of the reward and of the transition model. The main difference is that DRL quantifies the intrinsic stochasticity of the return, while in WQL the stochasticity refers to the uncertainty on the Q-function estimate which reduces as the number of updates increases, being a sample mean. We will reserve more space for this comparison in the final version.

3. (*Experiments*): we compared PDQN with two baselines: the standard Double DQN (DDQN) and Bootstrapped DQN (BDQN), meant to enforce exploration. See **General Note on Experiments** for details.

Reviewer #2

1. (*PDQN*) We are aware that the description of PDQN is synthetic, due to space constraints. We believe that the application to deep-RL should not be considered the main focus of the paper. As stated in **General Note on Experiments** our goal with PDQN is purely illustrative, showing the ability of WQL to be applied to continuous domains. More details about PDQN and the experimental setting are reported in Appendix C.2. We will insert the pseudocode of the PDQN algorithm in the final version. The update rules used in the case of a particle model for the Q-posteriors are reported in Table 1. The learning rate α_t can be set using for instance RMSProp. We stress that our theoretical findings (Section 5) are limited to the tabular case with no function approximation and, thus, are not applicable to PDQN. Extending the theory to function approximation is an appealing future research direction.

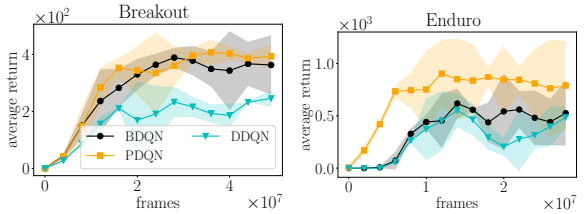
2. (*Experiments*) In PDQN the prior is the spread of the particles (see Appendix C.2.2). It determines the amount of exploration (similarly to BDQN) and how posteriors are built/updated. See also **General Note on Experiments**.

Reviewer #3

1. (*Proposition A.3*) We assumed that the particles are ordered as in Proposition A.1, we will state the assumption.

2. (*About the choice of Wasserstein*) The main reason why we chose the Wasserstein metric over other distributional distances (e.g., α-divergences) is that the Wasserstein distances are able to deal with deterministic distributions (α-divergences degenerate to infinity). This feature is important for us, as the Q-posteriors model a sample mean and, thus, their variances reduce as the number of samples increases, moving towards a deterministic distribution. This case is discussed in Proposition 3.1, in which the geometry of the support ($\mathbb{R}$) becomes essential in order to average deterministic distributions and recover the standard TD solution. This would not be possible if we employed α-divergences. Exploiting OT to account for the geometry of the state space is an appealing idea, although rather different from what we propose in our paper. In this regard, there are several works that exploit Lipschitz assumptions to incorporate the geometry of the MDP also using the Wasserstein (or Kantorovich) distances (e.g., [1]). We will explain, in the final version, the geometric motivations behind the choice of the Wasserstein distance and we will discuss in Section 7 the use of OT in RL.

General Note on Experiments We are aware that our experiments on Atari games are very limited. However, we believe that the focus of the paper is not the proposal of an effective deep-RL algorithm, but an algorithm that uses modern notions of Wasserstein barycenters, endowed with strong theoretical guarantees (PAC-MDP in average loss), which can be easily extended to function approximation, unlike other provably-efficient algorithms (e.g., MBIE, Delayed-QL). The choice of the Asterix game derives from the fact that, to see the advantages of PDQN, we need an environment in which exploration is essential. During the rebuttal period we run experiments on new environments: Breakout and Enduro (see figure). Unfortunately, due to our limited infrastructure, we could only show the average of 3 runs. Since in Breakout exploration is not an issue, the performance of PDQN is similar to BDQN but better than DDQN. However, in Enduro we can appreciate that PDQN learns substantially faster w.r.t. BDQN and DDQN, as exploration becomes relevant. We stress that all 3 algorithms use the same hyper-parameters (network, learning rate, replay buffer), which were not tuned. We will include these experiments in the final version.



[1] E. Rachelson and M. G. Lagoudakis. On the locality of action domination in sequential decision making. *ISAIM*, 2010.

[2] A. L. Strehl and M. L. Littman. An analysis of model-based interval estimation for markov decision processes. *Journal of Computer and System Sciences*, 74(8):1309–1331, 2008.