Anti-Exploration by Random Network Distillation

Alexander Nikulin, Vladislav Kurenkov, Denis Tarasov, Sergey Kolesnikov @ Tinkoff AI

Offline RL as Anti-Exploration Problem

Offline RL can be framed as an anti-exploration (Rezaeifar et al., 2022) problem. An online RL agent seeks to explore an uncertain in the state space, while an offline RL should avoid the uncertain in the action space. Such a framework allows unifying offline and online RL so that one method can be used in both settings, changing **only the sign of the exploration bonus**:

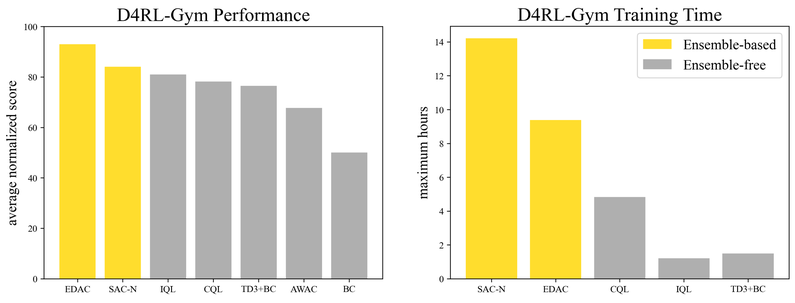
$$Q(s,a) = \mathbb{E}_D[r + \gamma \mathbb{E}_{a' \sim \pi(\cdot|s')}[Q(s',a') - \beta \cdot b(s',a')]]$$

Most of the recent successful uncertainty-based methods with ensembles, such as SAC-N, EDAC, MSG, and RORL, are implicitly or explicitly also anti-exploration approaches:

$$\mathbb{E}\left[\min_{j=1,\dots,N} Q_j(\mathbf{s},\mathbf{a})\right] \approx m(\mathbf{s},\mathbf{a}) - \Phi^{-1}\left(\frac{N-\frac{\pi}{8}}{N-\frac{\pi}{4}+1}\right) \cdot \sigma(\mathbf{s},\mathbf{a})$$

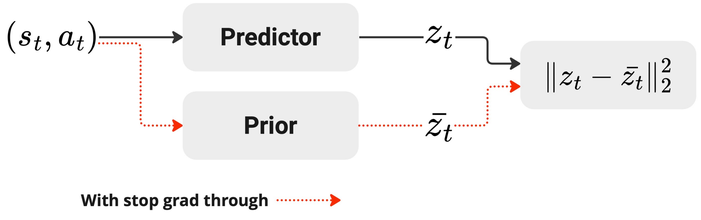
Ensembles are Effective but Slow

Although ensembles have state-of-the-art results on all D4RL datasets, **they require large Q-ensembles**, increasing computational and memory requirements. Can we do better?



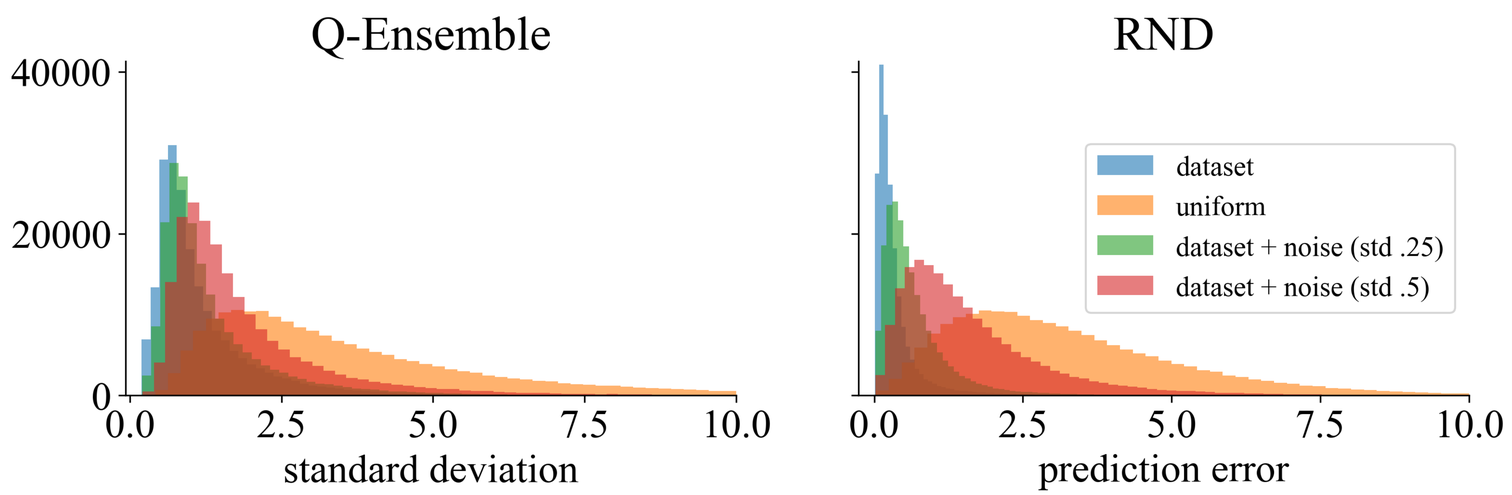
There is a Faster Alternative

We propose to use random network distillation (RND). It is a simple and fast competitor to ensembles for epistemic uncertainty estimation (Ciosek et al., 2019). Contrary to the Q-ensembles, it requires **only two additional networks**.



Random Network Distillation is Discriminative Enough

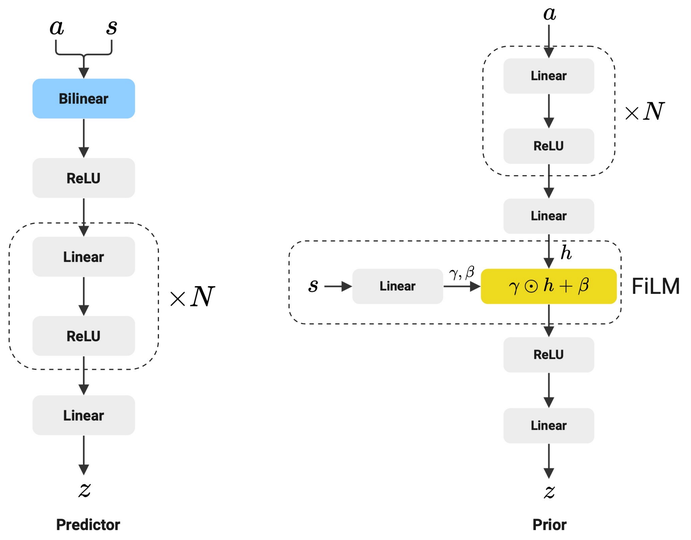
We show that, contrary to the previous results by Rezaeifar et al. (2022), RND is discriminative enough to be used as an uncertainty estimator to penalize out-of-distribution actions and is comparable to Q-ensembles.

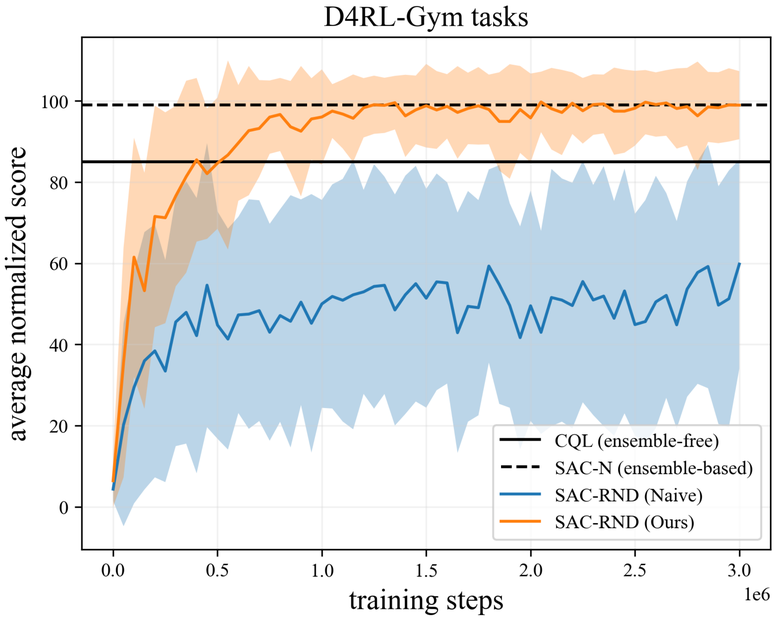


Multiplicative Conditioning is Necessary for Competitive Results

With simple state-action concatenation conditioning in the RND prior, it becomes infeasible for the actor to minimize the anti-exploration bonus effectively.

We explored many different ways of conditioning and demonstrated that **FiLM** significantly improves the ease of bonus minimization by the actor.





ICML
International Conference On Machine Learning