Behavior Preference Regression for Offline Reinforcement Learning

Padmanaba Srinivasan, William Knottenbelt

Department of Computing, Imperial College London
{ps3416, wjk}@imperial.ac.uk

Abstract

Offline reinforcement learning (RL) methods aim to learn optimal policies with access only to trajectories in a fixed dataset. Policy constraint methods formulate policy learning as an optimization problem that balances maximizing reward with minimizing deviation from the behavior policy. Closed form solutions to this problem can be derived as weighted behavioral cloning objectives that, in theory, must compute an intractable partition function. Reinforcement learning has gained popularity in language modeling to align models with human preferences; some recent works consider *paired* completions that are ranked by a preference model following which the likelihood of the preferred completion is directly increased. We adapt this approach of paired comparison. By reformulating the paired-sample optimization problem, we *fit* the maximum-mode of the Q function while maximizing *behavioral consistency* of policy actions. This yields our algorithm, Behavior Preference Regression for offline RL (BPR). We empirically evaluate BPR on the widely used D4RL Locomotion and Antmaze datasets, as well as the more challenging V-D4RL suite, which operates in image-based state spaces. BPR demonstrates state-of-the-art performance over all domains. Our on-policy experiments suggest that BPR takes advantage of the stability of on-policy value functions with minimal perceptible performance degradation on Locomotion datasets.

Introduction

As reinforcement learning (RL) sees increasing application in a variety of fields, from control (Razzaghi et al. 2022) to language modeling (Christiano et al. 2017), it has also become increasingly *data-hungry* (Shalev-Shwartz, Shamir, and Shammah 2017). The need to acquire data through *online* interaction can make deep reinforcement learning infeasible in many domains. In response, one direction of research develops *offline* RL algorithms that aim to learn from a static dataset of pre-collected interactions (Lange, Gabel, and Riedmiller 2012).

Standard off-policy algorithms can be directly applied on offline datasets (Haarnoja et al. 2018; Gulcehre et al. 2020), though in practice the combined effect of off-policy learning, bootstrapping, and function approximation (Sutton and Barto 2018) introduces extrapolation error. The resulting distribution shift between the learned policy and behavior policy can cause training instability and subsequent failure when deployed in the real environment (Fujimoto, Meger, and Precup 2019).

Offline RL algorithms address the challenges of offline off-policy evaluation in one of three ways: 1) incorporating pessimism into value estimation, 2) imposing policy constraints or 3) avoiding off-policy evaluation altogether by learning an on-policy value function. Pessimism offers performance guarantees (Jin, Yang, and Wang 2021), policy constraints may make better use of the representational power of neural networks (Geng et al. 2022) and learning on-policy values is more stable and avoids the overestimation and iterative exploitation associated with off-policy evaluation (Brandfonbrener et al. 2021).

Another approach to learning aims to align policy rollouts with human preferences (Akrour, Schoenauer, and Sebag 2011; Cheng et al. 2011; Christiano et al. 2017). Preference-based RL is popular in language modeling under the banner of RL from human feedback. Recent methods take models trained using supervised learning and finetune them using an offline dataset by directly increasing the likelihood of generating preferred sequences (Rafailov et al. 2024).

The principle of aligning policies with human preferences has been explored in offline RL (Kim et al. 2023; Rafailov et al. 2024; Hejna et al. 2023). While they aim to solve the same tasks, preference-based methods must either directly learn to generate aligned sequences (Kim et al. 2023) or must train a preference model (Rafailov et al. 2024; Hejna et al. 2023) on specially crafted datasets of human preferences. These methods typically eschew more traditional reward modeling (RM) and perform in-sample learning using pairs of trajectories.

Contributions Motivated by finetuning approaches to align language models (Rafailov et al. 2024; Gao et al. 2024), in this work we develop a policy objective for offline RL that directly learns the policy density: our algorithm performs Behavior Preference Regression for offline RL (BPR). We analyze BPR with respect to regularized value functions in the context of preference models to demonstrate theoretical performance improvement. Evaluation on D4RL (Fu et al. 2020) demonstrates that BPR achieves SOTA performance on Locomotion and Antmaze datasets. Additional

tests on the image-based V-D4RL (Lu et al. 2022) tasks reveal that BPR is able to transition across modalities to achieve high performance in non-proprioceptive domains. In experiments with on-policy value functions, BPR outperforms competing methods by a substantial margin on four of six datasets. By incorporating more expressive ensembles of value functions, BPR improves performance substantially on tasks that typically require trajectory stitching.

Related Work

Reinforcement learning aims to solve sequential decision-making tasks formulated as a Markov Decision Process (MDP), $\mathcal{M} = \{\mathcal{S}, \mathcal{A}, \mathcal{R}, P, p_0, \gamma\}$, where $\mathcal{S}$ denotes the state space, $\mathcal{A}$ the action space, $\mathcal{R}$ a scalar reward function, P the transition dynamics, p_0 the initial state distribution, and $\gamma \in [0, 1)$ the discount factor. The goal of RL is to learn an optimal policy that executes actions such that it maximizes the expected discounted reward; for any policy π we denote its return as $\eta(\pi) = \mathbb{E}_{\tau\sim\rho_\pi(\tau)}\left[\sum_{t=0}^{T}\gamma^t\mathcal{R}(s_t, a_t)\right]$ where $\rho_\pi(\tau) = p_0(s_0)\prod_{t=1}^{T}\pi(a_t|s_t)P(s_{t+1}|s_t, a_t)$ is a trajectory sampled under policy π (Sutton and Barto 2018).

Offline Reinforcement Learning

Offline RL methods aim to maximize sample efficiency and learn optimal policies given only a static dataset of interactions $\mathcal{D} = \{s, a, r, s\}_{n=1}^{N}$, which was produced by one or more unknown behavior policies of uncertain quality.

The tuples that form the dataset contain information that we are certain about. Actions beyond the support of the dataset are of unknown quality and lead to unknown trajectories. Generally, offline RL methods aim to train policies that maximize expected reward while remaining within the dataset support.

Off-Policy Methods A large body of offline RL methods adapt existing off-policy algorithms for the offline domain. Approaches can be classified into those that apply critic regularization to address overestimation and those that impose policy constraints to draw the current policy towards the dataset support.

Critic regularizers can explicitly reduce the values of OOD actions (Kumar et al. 2020; Kostrikov et al. 2021), thus shaping the Q function. This forces the policy to maximize Q values that are in-support. Regularizers can function implicitly by making use of the diversity-based-pessimism of large ensembles of value functions (An et al. 2021; Ghasemipour, Gu, and Nachum 2022). Ensembles condone some degree of OOD action selection which An et al. (2021) attribute to improving performance. Fu, Wu, and Boulet (2022) explore this further and find that relaxing constraints can improve performance in algorithms without large ensembles. Work by Ghasemipour, Gu, and Nachum (2022) suggests that large min-clipped ensembles may be redundant due to the collapse in independence of ensemble members.

Policy constraints aim to directly confine the actor to select in-support actions. These are typically formulated explicitly as divergence penalties (Wu, Tucker, and Nachum 2019; Fujimoto and Gu 2021), implicitly through weighted behavioral cloning (BC) (Wang et al. 2020; Peng et al. 2019; Nair et al. 2020) or by architecturally limiting the exploration afforded to the policy (Fujimoto, Meger, and Precup 2019).

On-Policy Methods Brandfonbrener et al. (2021) recognize off-policy evaluation as a source of instability in offline RL and instead learn an on-policy (Onestep) value function. The policy learned using this value function outperforms those learned via behavioral cloning and some offline off-policy methods. On-policy learning is extended by Kostrikov, Nair, and Levine (2021) and Garg et al. (2023) who attempt to approximate the in-sample maximum return by dataset trajectories which they use to train a weighted BC policy. Zhuang et al. (2023) adapt online, on-policy PPO (Schulman et al. 2017) for the offline setting and develop an algorithm that uses offline datasets with periodic online evaluation. This is not a fully offline RL algorithm and their own experiments show that without online evaluation to enable policy replacement, performance will degrade.

Preference-Based Reinforcement Learning

Building on the ideas of Akrour, Schoenauer, and Sebag (2011) and Cheng et al. (2011), Christiano et al. (2017) suggest using preference-annotated data as reward signals to train language models that are better aligned with human values. Subsequent work has developed learning from preferences further (Kaufmann et al. 2023) with the notable *Direct Preference Optimization* (DPO) (Rafailov et al. 2024) which finetunes a maximum likelihood trained policy on an offline dataset of paired preference annotated data by directly optimizing policy density as a proxy for the reward function.

In continuous-control offline RL, Kim et al. (2023) train a trajectory-producing policy on non-Markovian, preference-based rewards. An et al. (2023) use a preference labeled dataset to train a preference model that is subsequently used to label preferred trajectories in an unlabeled dataset used for policy training. Using preference datasets, Hejna et al. (2023) directly train a policy (similar to DPO) as an optimal advantage function using preference data. Common themes of preference-based offline RL methods are the eschewing of traditional rewards for human-annotated data, and the requirement trajectories to be *paired* for preference learning which does not allow evaluation of OOD actions.

Behavior Preference Regression

We consider the general, reverse KL-constrained problem:

$$\pi_{t+1} = \arg\max_{\pi\in\Pi}\mathbb{E}_{s\sim\mathcal{D},a\sim\pi}[f(s,a) - \lambda D_{\text{KL}}(\pi(\cdot|s)||\pi_{\text{ref}}(\cdot|s))], \quad (1)$$

where $\lambda \geq 0$ controls the tradeoff between remaining close to a distribution π_{ref} and maximizing some function $f(\cdot, \cdot)$.

The closed form solution to the optimization problem has been previously derived (Ziebart et al. 2008; Grünwald and

Dawid 2004):

$$\pi_{t+1} = \pi_{\text{ref}}(a|s) \exp(\frac{1}{\lambda} f(s,a)) \frac{1}{Z(s)} \tag{2}$$

$$Z(s) = \int_{a \in \mathcal{A}} \pi_{\text{ref}}(a|s) \exp(\frac{1}{\lambda} f(s,a))\, da, \tag{3}$$

where $Z(s)$ is the partition function.

Using the *DPO trick* (Rafailov et al. 2024), we can rearrange Equation 2 as:

$$f(s,a) = \lambda \left(\log Z(s) + \log \frac{\pi_{t+1}(a|s)}{\pi_{\text{ref}}(a|s)} \right), \tag{4}$$

following which using ranked, paired samples where $a_1 \succ a_2$ we can write:

$$f(s,a_1) - f(s,a_2) = \lambda \left(\log \frac{\pi_{t+1}(a_1|s)}{\pi_{\text{ref}}(a_1|s)} - \log \frac{\pi_{t+1}(a_2|s)}{\pi_{\text{ref}}(a_2|s)} \right), \tag{5}$$

which conveniently cancels out the partition function.

DPO takes the binary preference $a_1 \succ a_2$ and passes the RHS through a Bradley-Terry preference model (Bradley and Terry 1952) to optimize for a_1. Consequently, DPO fails to capture *how much more* a_1 is preferred to a_2. Gao et al. (2024) aim to directly learn the relative difference by solving the regression problem:

$$\big[(f(s,a_1) - f(s,a_2)) - \lambda \left(\log \frac{\pi_{t+1}(a_1|s)}{\pi_{\text{ref}}(a_1|s)} - \log \frac{\pi_{t+1}(a_2|s)}{\pi_{\text{ref}}(a_2|s)} \right) \big]^2. \tag{6}$$

In this work, we focus on learning a policy by solving this relative regression problem.

What do we *Prefer* in Offline RL?

Most policy constraint formulations typically choose $f(s,\cdot) = Q(s,\cdot)$ and $\pi_{\text{ref}}(\cdot|s) = \hat{\pi}_\beta(\cdot|s)$ where $\hat{\pi}_\beta$ is an empirical behavior policy. This follows the principle of maximizing reward while satisfying some constraint that must be carefully balanced by tuning λ to curb the distribution shift (Brandfonbrener et al. 2021).

We propose an alternative optimization: we **maximize behavioral consistency** and **reverse KL fit the (maximum) mode of the Q function** – in preference terms, we fit a distribution of high-reward actions and regress toward actions with high likelihood under the behavior policy.

Selecting π_{ref} Soft Q-learning (Haarnoja et al. 2018) trains a maximum entropy Q function that can be written as an energy-based model (EBM) (Goodfellow, Bengio, and Courville 2016). We formulate $\pi_{\text{ref}}(a|s) = \frac{\exp(Q(s,a))}{Z_Q(s)}$ where $Z_Q(s) = \int_{\mathcal{A}} Q(s,a)\, da$ is the partition function, which subsequently cancels out in the RHS of Equation 6. This allows us to directly optimize the soft actor–critic (SAC) policy objective (Haarnoja et al. 2018) without resorting to approximations of the entropy through a `tanh`-transformed Gaussian.

Selecting $f(\cdot,\cdot)$ The true behavior policy is unknown and so we must make an empirical approximation. Prior methods typically learn explicit policies using behavioral cloning (Kostrikov et al. 2021; Wu, Tucker, and Nachum 2019; Zhuang et al. 2023). This can be limiting, as the number of behavior policy modes must be known beforehand. Implicit policies offer more flexible behavior models (Florence et al. 2022). We train an implicit behavior policy $\hat{\pi}_\beta$ as an EBM that learns an energy function $E(s,a) \in \mathbb{R}$. We recover an estimate of the explicit behavior policy using the Boltzmann distribution: $\hat{\pi}_\beta(a|s) = \frac{\exp(-E(s,a))}{Z_E(s)}$ where $Z_E(s) = \int_{\mathcal{A}} \exp(-E(s,a))\, da$ is the EBM partition function.

Fortunately, $Z_E(s)$ also cancels out when using $f(s,\cdot) = \log \hat{\pi}_\beta(\cdot|s)$ in the LHS of Equation 6 and we only need to compute $E(s,a_1)$ and $E(s,a_2)$. Using an EBM behavior policy, we make no *inductive bias* with respect to the (multi)modality of the true behavior policy.

Combining everything, our policy optimization objective is:

$$[(E(s,a_2)) - E(s,a_1)) - \lambda \left(\log \frac{\pi_{t+1}(a_1|s)}{\exp(Q(s,a_1))} - \log \frac{\pi_{t+1}(a_2|s)}{\exp(Q(s,a_2))} \right)]^2. \tag{7}$$

Interpretation Learning π^* requires a policy to select in-sample actions that also maximize expected reward. By selecting the regression target to be the difference $\log \hat{\pi}_\beta(a_1|s) - \log \hat{\pi}_\beta(a_2|s)$, we treat the behavior EBM as an expert preference model that communicates by how much $a_1 \succ a_2$. This differs from previous preference-based offline RL formulations that evaluate the preference by comparing discounted rewards over entire trajectories (produced by the behavior policy) for a pair of actions. Such reward-based preference learning has been shown to be inconsistent with human-preference labels (Knox et al. 2022). Placing a support constraint on the policy towards high-reward modes in the soft Q function and combining this with off-policy evaluation offers a far more flexible approach without the need for human-labeled preference datasets. Most importantly, we **never** need to compute any partition function $Z(s), Z_Q(s)$ or $Z_E(s)$ – past work has found that approximating partition functions, though technically correct, is deleterious to performance (Nair et al. 2020).

Self-Play

Let μ_1, μ_2 be the sampling distributions for a_1 and a_2, respectively. Offline preference-based methods use datasets that contain previously evaluated pairs of completions sampled from π_β (Kim et al. 2023; Rafailov et al. 2024; Hejna et al. 2023). In standard offline settings, samples are drawn from $\mathcal{D}$ or π and in the paired setting this equates to using $\mu_1 = \pi_\beta = \mathcal{D}$ and $\mu_2 = \pi$ (reference sampling). Recently, Swamy et al. (2024) prove that performing self-play with multiple samples drawn from π itself results in stable learning with strong theoretical guarantees – this involves sampling a pair of actions from the current policy and querying a learned preference/reward model to optimize Equation 2. We use self-play to sample actions for policy optimization,

hence $\mu_1 = \mu_2 = \pi$. We compare reference sampling and self-play schemes in a toy bandit example in the Appendix.

Analysis

Rearranging Equation 5 and inserting $\pi_{\text{ref}}(\cdot|s) = \frac{\exp(Q(s,\cdot))}{Z_Q(s)}$ and $f(s,\cdot) = \log \pi_\beta(\cdot|s)$, we obtain:

$$\begin{aligned}(Q(s,a_1) + \frac{1}{\lambda}\log \pi_\beta(a_1|s)) - \\ (Q(s,a_2) + \frac{1}{\lambda}\log \pi_\beta(a_2|s)) \\ = \log \pi_{t+1}(a_1|s) - \log \pi_{t+1}(a_2|s). \quad (8)\end{aligned}$$

We explicitly cancel $Z_Q(s)$ but leave $Z_E(s)$ unfactorized for clarity.

We define $\tilde{Q}(s,a) \triangleq Q(s,a) + \frac{1}{\lambda}\log \hat{\pi}_\beta(a|s)$ and notice that this is a variation of an implicit Q function popular in *online* RL (Vieillard et al. 2021; Peters, Mulling, and Altun 2010) and is exactly the Q function formulation used by Fisher-BRC when using $\lambda = 1.0$ (Kostrikov et al. 2021). We subsequently interpret that our policy regression objective is equivalent to fitting the policy to the implicit Q function.

We rewrite the LHS of Equation 8 as a *soft* preference function:

$$\mathtt{P}(s,a_1,a_2) \triangleq \tilde{Q}(s,a_1) - \tilde{Q}(s,a_2). \quad (9)$$

Assumption 1 *(Tuned Preference Function)*

$$\begin{aligned}P(s,a_1,a_2) \geq 0 \quad \forall a_1,a_2 \in \mathcal{A} \\ \textit{when} \quad \pi_\beta(a_1|s) \geq \pi_\beta(a_2|s).\end{aligned}$$

We assume that any action a_1 with a higher likelihood under the behavior policy than a_2 is preferred. In practice, this can be satisfied by tuning λ.

For any policy π, recall its return is given by $\eta(\pi) = \mathbb{E}_{\tau\sim\rho_\pi(\tau)}\left[\sum_{t=0}^{T}\gamma^t \mathcal{R}(s_t,a_t)\right]$. The behavior policy used to produce the dataset is π_β and let the policy learned by optimizing using $\tilde{Q}(s,a)$ be $\tilde{\pi}$ (i.e. the policy that maximizes soft preferences).

Proposition 1 *(Perfect Preference Model) If the preference function $P(s,a_1,a_2)$ is perfect i.e. $\tilde{Q}^* = Q^* + \pi_\beta$ is accurate, then the deterministic policies π_β and $\tilde{\pi}$ satisfy:*

$$\begin{aligned}\eta(\tilde{\pi}) - \eta(\pi_\beta) \\ \approx \mathbb{E}_{s\sim\mathcal{D}}\left[\tilde{Q}^*(s,\tilde{\pi}(s)) - \tilde{Q}^*(s,\pi_\beta(s))\right] \geq 0 \quad (10)\end{aligned}$$

In practice, estimation is noisy. For $\tilde{Q}$, this comes from two sources: errors are present in both Q function and behavior policy estimates. EBM approximation error has been studied by Florence et al. (2022) (Theorem 2) who prove that a Lipschitz-continuous EBM policy can exhibit arbitrarily small error.

The total variational distance between two value functions Q_1, Q_2 is: $D_{\text{TV}}(Q_1,Q_2) = \max_{s\in\mathcal{S}}|Q_1(s,\pi(s)) - Q_2(s,\pi(s))|$.

Algorithm 1: Policy improvement step. Comment NG denotes steps where gradients do not have to be computed.

Require: Offline dataset $\mathcal{D}$, pretrained EBM $E(\cdot,\cdot)$, training steps N
Output: Trained policy π

```
Let t = 0.
for t = 1 to N do
    Sample (s, a, r, s') ~ D
    Sample a_1, a_2 ~ π        # NG
    Compute log π(a_1|s), log π(a_2|s)
    Compute E(s, a_1) and E(s, a_2)    # NG
    Compute Q(s, a_1) and Q(s, a_2)    # NG
    Update π using Equation 7.
    # Update critics
end for
return π
```

Proposition 2 *(Noisy Preference Model) Consider the case where $\hat{\pi}_\beta$ and Q^* contain errors and produce the noisy $\tilde{Q}^-$. Then $\forall\tilde{Q}^-$ where $D_{TV}(\tilde{Q}^-(s,\tilde{\pi}(s)), Q^*(s,\tilde{\pi}(s))) \leq \tilde{\epsilon}$ and $D_{TV}(\tilde{Q}^-(s,\pi_\beta(s)), Q^*(s,\pi_\beta(s))) \leq \epsilon$ the following holds:*

$$\begin{aligned}\eta(\tilde{\pi}) - \eta(\pi_\beta) \\ \leq \mathbb{E}_{s\sim\mathcal{D}}\left[\tilde{Q}^-(s,\tilde{\pi}(s)) - \tilde{Q}^-(s,\pi_\beta(s))\right] \\ + 2\rho_{max}(\tilde{\epsilon}+\epsilon) \quad (11) \\ \textit{where} \quad \rho_{max} = \sup\{\rho_{\pi_\beta}(s), s\in\mathcal{S}\}.\end{aligned}$$

We defer proofs to the Appendix.

The first term after the inequality is non-negative under Assumption 1 and the second term is present due to the modeling error of the estimated Q function and behavior policy. This can be reduced by using a more accurate function approximator.

Implementation

Our actor–critic implementation follows a standard implementation of SAC (Haarnoja et al. 2018) with modifications to the policy improvement step. We illustrate our policy improvement in Algorithm 1 and provide additional implementation details in the Appendix.

The EBM approximation of π_β is trained prior to the main actor–critic training phase. We follow design decisions detailed in Florence et al. (2022), using spectral normalization (Miyato et al. 2018) and deep networks.

Summary of Hyperparameters In addition to the standard hyperparameters of SAC (clipped double-Q learning (Fujimoto, Hoof, and Meger 2018), entropy regularized off-policy Q functions), our algorithm introduces the hyperparameter λ, which controls the tradeoff between the KL constraint and maximizing behavioral consistency.

In general, we find that simply **using $\lambda = 1.0$ works well across all tasks**; our primary results use this hyperparameter value and we perform ablations to evaluate sensitivity in our experiments.

Experiments

In this section, we evaluate empirically BPR and aim to answer the following questions:

- How well does BPR perform compared to state-of-the-art offline RL methods?
- Does BPR perform well in tasks with visual state spaces?
- Can Onestep-trained policies compete with off-policy offline RL?
- How sensitive is BPR to values of λ?

Experimental Setup In all BPR experiments, we report the normalized mean score with standard deviation on five seeds over 100 evaluations in Antmaze tasks and 10 in others. All scores are reported using the policy from the final checkpoint.

Baselines We compare results against the following, well-known baselines: CQL (Kumar et al. 2020), IQL (Kostrikov, Nair, and Levine 2021) and TD3+BC (Fujimoto and Gu 2021). We also include the recent offline RL algorithms: ReBRAC (Tarasov et al. 2023), XQL (Garg et al. 2023) and Diff-QL (Wang, Hunt, and Zhou 2022) (which replaces a Gaussian/deterministic policy with a Diffusion policy (Ho, Jain, and Abbeel 2020)). Of the latter three methods, both ReBRAC and XQL tune hyperparameters extensively for each dataset. In contrast, the older baselines, Diff-QL and our BPR find hyperparameters that generalize well across like-tasks (i.e. the same hyperparameters for all Locomotion tasks etc.).

For a more comprehensive comparison, we also include the preference-based offline RL methods: PT (Kim et al. 2023), OPPO (Kang et al. 2023) and DPPO (An et al. 2023).

D4RL

We evaluate BPR on D4RL Locomotion and Antmaze datasets (Fu et al. 2020).

Locomotion The Locomotion datasets offer varying degrees of suboptimality, using mixtures of highly suboptimal trajectories (`-replay`) and optimal ones (`-expert`). Table 1 shows BPR's Locomotion scores. In general, all methods recover near-expert performance on any `expert` datasets. BPR greatly outscores all older baselines as well as preference-based algorithms. ReBRAC is highly tuned for each dataset and BPR, for the most part, scores similarly except for `hc-m` (where ReBRAC scores higher), and `w-m` and `w-m-r`, where BPR outperforms ReBRAC by a substantial margin.

Antmaze The Antmaze tasks are characterized by sparse reward schemes and suboptimal trajectories which necessitates off-policy evaluation (or IQL/XQL in-sample max estimation) to perform well. In the smaller mazes, BPR, ReBRAC and XQL perform similarly, though BPR is able to sustain high performance as the maze grows. Preference-based PT does not perform well in larger mazes.

V-D4RL

Most offline RL algorithms typically limit their evaluation to proprioceptive state spaces. V-D4RL (Lu et al. 2022) is a benchmarking suite that evaluates offline RL algorithms in visual state spaces on continuous control tasks with mixtures of trajectories similar to those found in D4RL Locomotion and based on the DMC environments (Yarats et al. 2021).

The V-D4RL paper provides scores for CQL, and behavioral cloning (BC) policies, as well as LOMPO (Rafailov et al. 2021) and a variant of DrQ (Yarats et al. 2021) with a behavioral cloning constraint. LOMPO and DrQ are designed specifically to learn from visual state spaces. We also include results for ReBRAC, which is again tuned for each dataset. We use V-D4RL environments without distractors following Tarasov et al. (2023).

We present V-D4RL results in Table 3. Generally, BC outperforms CQL – the standard offline RL baseline. ReBRAC, with the help of tuning, is able to slightly outperform BC. BPR consistently outperforms the image-adapted LOMPO and DrQ+BC, trading blows with ReBRAC on `walker-walk` and `cheetah-run` datasets and keeps pace with BC on the more difficult `humanoid-walk` tasks.

Onestep Experiments

Off-policy evaluation can lead to querying and backing up of overestimated OOD actions that the policy can exploit, leading to instability. Onestep value functions are highly stable due to their on-policy nature (Brandfonbrener et al. 2021) and recent work by Eysenbach et al. (2023) shows equivalence between Onestep values and CQL-style critic regularization.

We evaluate how well BPR with a Onestep value function performs compared to the original Onestep RL (O-RL) algorithm (Brandfonbrener et al. 2021). We also include Locomotion results from CFPI (Li et al. 2023), which uses a first-order Taylor approximation as a linear approximation of the Q function, and trains a Onestep value function using distributional critics (Dabney et al. 2018a,b).

We report results on non-expert Locomotion datasets and the `medium` and `large` Antmaze datasets in Table 4. Both Onestep RL and CFPI perform similarly on Locomotion tasks. BPR matches their performance on two tasks and outperforms both by a large margin on four out of six Locomotion tasks.

Onestep RL performs poorly on the `medium` and `large` Antmaze tasks. In contrast, BPR is able to make significant progress in all these sparse reward tasks, falling slightly short of off-policy CQL (see Table 2).

Suboptimality in D4RL The similarity in performance between Onestep BPR and off-policy BPR in Locomotion tasks suggests that trajectories in these datasets may not be as suboptimal as originally thought (Fu et al. 2020). This explains the recent saturation in performance on Locomotion (Tarasov et al. 2023). Antmaze, while challengingly suboptimal, may be a poor evaluator of generalization (Rafailov et al. 2024). The performance of Onestep BPR indicates that this may be a pragmatic variant to select for application due to its improved stability.

Dataset	CQL	IQL	TD3+BC	ReBRAC	XQL	Diff-QL	PT	OPPO	DPPO	**BPR (ours)**
hc-m	44.0	47.4	48.3	**65.6**	48.3	51.1	-	43.4	-	53.7 ± 1.4
hp-m	58.5	66.3	59.3	**102.0**	74.2	90.5	-	86.3	-	101.3 ± 1.1
w-m	72.5	78.3	83.7	82.5	84.2	87.0	-	85.0	-	**91.1 ± 3.7**
hc-m-r	45.5	42.2	44.6	**51.0**	45.2	47.8	-	39.8	40.8	50.9 ± 0.6
hp-m-r	95.0	94.7	60.9	98.1	100.7	101.3	84.5	88.9	73.2	**102.0 ± 4.9**
w-m-r	77.2	73.9	81.8	77.3	82.2	95.5	71.3	71.7	50.9	**97.4 ± 2.7**
hc-m-e	91.6	86.7	90.7	101.1	94.2	96.8	-	89.6	92.6	**103.8 ± 4.3**
h-m-e	105.4	91.5	98.0	107.0	**111.2**	111.1	69.0	108.0	107.2	110.9 ± 5.2
w-m-e	108.8	109.6	110.1	111.6	**112.7**	110.1	110.1	105.0	108.6	110.8 ± 0.2

Table 1: Normalized scores on D4RL Gym Locomotion datasets. All scores are taken from their respective original papers. hc, hp and w refer to halfcheetah, hopper and walker2d environments, respectively. Methods are grouped by: older baselines, newer offline RL baselines, preference-based offline RL methods followed by BPR. For XQL, we use the per-dataset tuned variant's scores. We report SD for BPR and **bold** the top score and underline BPR scores when within 1 SD of the best.

Dataset	CQL	IQL	TD3+BC	ReBRAC	XQL	Diff-QL	PT	**BPR (ours)**
-umaze	74.0	87.5	78.6	**97.8**	93.8	93.4	-	95.6 ± 1.0
-umaze-d	84.0	62.2	71.4	88.3	82.0	66.2	-	**89.1 ± 1.1**
-medium-p	61.2	71.2	10.6	84.0	76.0	76.6	70.1	**86.7 ± 3.7**
-medium-d	53.7	70.0	3.0	76.3	73.6	78.6	65.3	**82.9 ± 7.8**
-large-p	15.8	39.6	0.2	60.4	46.5	46.4	42.4	**70.3 ± 8.3**
-large-d	14.9	47.5	0.0	54.4	49.0	56.6	19.6	**72.1 ± 5.1**

Table 2: Normalized scores on D4RL Antmaze datasets. Methods are grouped by: older baselines, newer RM RL baselines, preference-based offline RL methods followed by BPR. For XQL, we use the per-dataset tuned variant's scores. We report SD for BPR and **bold** the top score and underline BPR scores when within 1 SD of the best.

Dataset	BC	CQL	ReBRAC	LOMPO	DrQ+BC	**BPR (ours)**
ww-mixed	16.5 ± 4.3	11.4 ± 12.4	41.6 ± 8.0	34.7 ± 19.7	28.7 ± 6.9	**45.0 ± 11.2**
ww-medium	40.9 ± 3.1	14.8 ± 16.1	**52.5 ± 3.2**	43.9 ± 11.1	46.8 ± 2.3	50.7 ± 4.1
ww-medexp	47.7 ± 3.9	56.4 ± 38.4	92.7 ± 1.3	39.2 ± 19.5	86.4 ± 5.6	**97.4 ± 1.9**
cr-mixed	25.0 ± 3.6	10.7 ± 12.8	**46.8 ± 0.7**	36.3 ± 15.6	44.8 ± 3.6	45.0 ± 3.1
cr-medium	51.6 ± 1.4	40.9 ± 5.1	**58.3 ± 11.7**	16.4 ± 18.3	50.6 ± 8.2	55.3 ± 1.2
cr-medexp	57.5 ± 6.3	20.9 ± 5.5	58.3 ± 11.7	11.9 ± 1.9	50.6 ± 8.2	**62.7 ± 8.5**
hw-mixed	**18.8 ± 4.2**	0.1 ± 0.0	16.0 ± 2.7	0.2 ± 0.0	15.9 ± 3.8	18.3 ± 1.9
hw-medium	**13.5 ± 4.1**	0.1 ± 0.0	9.0 ± 2.3	0.1 ± 0.0	6.2 ± 2.4	9.0 ± 0.8
hw-medexp	**17.2 ± 4.7**	0.1 ± 0.0	7.8 ± 2.4	0.2 ± 0.0	7.0 ± 2.3	13.3 ± 4.4

Table 3: Normalized scores on V-D4RL tasks. ww, cr and hw refer to walker-walk, cheetah-run and humanoid-walk environments, respectively. Methods are grouped by: BC, offline RL baselines, RL algorithms adapted for visual state spaces followed by BPR. We report 1 SD for all methods and **bold** the top score and underline BPR scores when within 1 SD of the best.

Dataset	O-RL	CFPI	**Onestep BPR**
hc-m	**55.6**	51.1	52.0 ± 0.8
hp-m	83.3	86.8	**96.4 ± 0.4**
w-m	85.6	88.3	**89.7 ± 1.3**
hc-m-r	41.4	44.5	**51.0 ± 0.4**
h-m-r	71.0	93.6	**99.1 ± 2.3**
w-m-r	71.6	78.2	**92.0 ± 0.8**
amaze-m-p	0.3	-	**52.7 ± 10.3**
amaze-m-d	0.0	-	**40.0 ± 7.8**
amaze-l-p	0.0	-	**10.4 ± 2.9**
amaze-l-d	0.0	-	**12.7 ± 1.6**

Table 4: Scores for Onestep BPR with Onestep RL and Onestep CFPI. We evaluate on non-expert Locomotion and `medium` and `large` Antmaze (`amaze`) datasets. The authors of CFPI do not report Onestep results for Antmaze. We report 1 SD for BPR and **bold** the top score and underline BPR scores when within 1 SD of the best.

More Expressive Onestep Value Functions O-RL uses a single Q function and samples actions to estimate state-value to compute advantage. CFPI trains two distributional critics and ses the min-clipped value estimate during bootstrapping. Onestep BPR trains two regular, min-clipped critics. Diversity can collapse in ensembles with shared targets. We investigate whether diversity at the cost of pessimism can improve performance; we experiment with Onestep, independent 4-critic ensembles to estimate the Q value lower confidence bound (Ghasemipour, Gu, and Nachum 2022):

$$Q_{\text{LCB}}(s,a) = \mathbb{E}^{\text{ens}}\left[Q_i(s,a)\right] - \omega \mathbb{V}^{\text{ens}}\left[Q_i(s,a)\right], \quad (12)$$

where $\mathbb{E}^{\text{ens}}$ and $\mathbb{V}^{\text{ens}}$ indicate mean and variance over the ensemble of Q functions and ω is a parameter that controls the degree of pessimism. We use $\omega = 2.0$ in all experiments.

Compared to Onestep BPR, Ensemble BPR sees performance improvements of **at least 10 points on each dataset** on the `medium` and `large` Antmaze datasets. Detailed per-dataset scores and implementation information can be found in the Appendix.

Ablations

Recall that λ controls the tradeoff between maximizing behavioral consistency and fitting the Q function in Equation 1. We examine sensitivity to λ for off-policy BPR in a series of ablation experiments in the D4RL Locomotion tasks.

Sensitivity to λ varies between datasets, with little performance variation on `halfcheetah-medium` and `halfcheetah-medium-replay`. In other datasets, using $\lambda = 0.5$ or $\lambda = 2.0$ sees performance decline. Our choice of $\lambda = 1.0$ generalizes well over all datasets and usually outperforms $\lambda = 1.5$. We provide detailed ablation results in the Appendix.

Discussion

Performance Our key contribution in this work is the development of a policy objective that reduces policy improvement to a regression problem. Off-policy BPR results in D4RL Locomotion datasets are on par with current SOTA and BPR **outperforms RL baselines in 5 out of 6 Antmaze datasets** and **6 out of 9 V-D4RL datasets**. Our Onestep experiments show that Onestep BPR outperforms Onestep RL in **9 out of 10 tasks** and CFPI in all Locomotion tasks. BPR requires **minimal tuning** to achieve high performance – **all our results are produced using $\lambda = 1.0$**.

Density Estimation Employing estimates of the behavior policy is common in many offline RL algorithms. Most prior works use explicit density estimates using Gaussian policies, mixture density networks (Bishop 1994) or VAEs (Kingma and Welling 2013). If the modality of the behavior policy is known, the first two methods can be used in BPR. VAEs are unsuitable as density estimation requires sampling.

The function $f(\cdot,\cdot)$ does not need to be a density estimate. Another natural choice for $f(\cdot,\cdot)$ is a discriminator (Goodfellow et al. 2014) that replaces a density estimate with an adversarial critic trained concurrently. This offers more choice of the exact f-divergence to minimize at the cost of increased training instability (Jolicoeur-Martineau 2020).

Critic Ensembles Our ensemble experiments imply that Onestep-trained policies might perform better than prior work reports. The *optimistic pessimism* of Q_{LCB} ensembles could enable algorithms to learn better policies while still enjoying the stability of on-policy evaluation.

Limitations EBMs can be difficult and computationally expensive to train. As a consequence of the Manifold hypothesis, they may also generalize poorly (Bengio, Courville, and Vincent 2013), though all models capable of multimodal learning suffer from their own slew of problems (Goodfellow, Bengio, and Courville 2016). Advancements in methodology have improved the stability of training and quality of models (Du and Mordatch 2019). Both prior work (Florence et al. 2022) and the results of our experiments suggest that EBMs are well-suited for offline RL.

Conclusion

In this paper, we introduce Behavior Preference Regression (BPR). Our method formulates a reframed, paired-sample policy objective that directly trains a policy likelihood to be behaviorally consistent and maximize reward, using least-squares regression. Though our method is motivated by fine-tuning approaches in language models, it is extensible to offline RL. We validate our algorithm on datasets with a variety of task types and reward schemes that offer both proprioceptive and image-based state spaces. BPR consistently outperforms prior RM-based approaches and preference-based ones by a substantial margin.

Additional experiments evaluating Onestep BPR demonstrate that our algorithm can learn policies that outperform previous Onestep methods. Furthermore, with more expressive Onestep value functions, BPR makes headway on the challenging Antmaze tasks that typically demand off-policy evaluation.

Future work should further review the viability of Onestep ensembles and look to adapt paired completion approaches for offline continuous control.

References

Akrour, R.; Schoenauer, M.; and Sebag, M. 2011. Preference-based policy learning. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2011, Athens, Greece, September 5-9, 2011. Proceedings, Part I 11*, 12–27. Springer. ISBN 3642237797.

An, G.; Lee, J.; Zuo, X.; Kosaka, N.; Kim, K.-M.; and Song, H. O. 2023. Direct preference-based policy optimization without reward modeling. *Advances in Neural Information Processing Systems*, 36: 70247–70266.

An, G.; Moon, S.; Kim, J.-H.; and Song, H. O. 2021. Uncertainty-based offline reinforcement learning with diversified Q-ensemble. *Advances in Neural Information Processing Systems*, 34: 7436–7447.

Bengio, Y.; Courville, A.; and Vincent, P. 2013. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8): 1798–1828.

Bishop, C. M. 1994. Mixture density networks.

Bradley, R. A.; and Terry, M. E. 1952. Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, 39(3/4): 324–345.

Brandfonbrener, D.; Whitney, W.; Ranganath, R.; and Bruna, J. 2021. Offline RL without off-policy evaluation. *Advances in Neural Information Processing Systems*, 34: 4933–4946.

Cheng, W.; Fürnkranz, J.; Hüllermeier, E.; and Park, S.-H. 2011. Preference-based policy iteration: Leveraging preference learning for reinforcement learning. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2011, Athens, Greece, September 5-9, 2011. Proceedings, Part I 11*, 312–327. Springer. ISBN 3642237797.

Christiano, P. F.; Leike, J.; Brown, T.; Martic, M.; Legg, S.; and Amodei, D. 2017. Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems*, 30.

Dabney, W.; Ostrovski, G.; Silver, D.; and Munos, R. 2018a. Implicit quantile networks for distributional reinforcement learning. In *International Conference on Machine Learning*, 1096–1105. PMLR. ISBN 2640-3498.

Dabney, W.; Rowland, M.; Bellemare, M.; and Munos, R. 2018b. Distributional reinforcement learning with quantile regression. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32. ISBN 2374-3468.

Du, Y.; and Mordatch, I. 2019. Implicit generation and modeling with energy based models. *Advances in Neural Information Processing Systems*, 32.

Eysenbach, B.; Geist, M.; Levine, S.; and Salakhutdinov, R. 2023. A connection between one-step RL and critic regularization in reinforcement learning. In *International Conference on Machine Learning*, 9485–9507. PMLR. ISBN 2640-3498.

Florence, P.; Lynch, C.; Zeng, A.; Ramirez, O. A.; Wahid, A.; Downs, L.; Wong, A.; Lee, J.; Mordatch, I.; and Tompson, J. 2022. Implicit behavioral cloning. In *Conference on Robot Learning*, 158–168. PMLR. ISBN 2640-3498.

Fu, J.; Kumar, A.; Nachum, O.; Tucker, G.; and Levine, S. 2020. D4RL: Datasets for deep data-driven reinforcement learning. *arXiv preprint arXiv:2004.07219*.

Fu, Y.; Wu, D.; and Boulet, B. 2022. A closer look at offline RL agents. *Advances in Neural Information Processing Systems*, 35: 8591–8604.

Fujimoto, S.; and Gu, S. S. 2021. A minimalist approach to offline reinforcement learning. *Advances in Neural Information Processing Systems*, 34: 20132–20145.

Fujimoto, S.; Hoof, H.; and Meger, D. 2018. Addressing function approximation error in actor-critic methods. In *International Conference on Machine Learning*, 1587–1596. PMLR. ISBN 2640-3498.

Fujimoto, S.; Meger, D.; and Precup, D. 2019. Off-policy deep reinforcement learning without exploration. In *International Conference on Machine Learning*, 2052–2062. PMLR. ISBN 2640-3498.

Gao, Z.; Chang, J. D.; Zhan, W.; Oertell, O.; Swamy, G.; Brantley, K.; Joachims, T.; Bagnell, J. A.; Lee, J. D.; and Sun, W. 2024. REBEL: Reinforcement Learning via Regressing Relative Rewards. *arXiv preprint arXiv:2404.16767*.

Garg, D.; Hejna, J.; Geist, M.; and Ermon, S. 2023. Extreme Q-learning: Maxent RL without entropy. *arXiv preprint arXiv:2301.02328*.

Geng, X.; Li, K.; Gupta, A.; Kumar, A.; and Levine, S. 2022. Effective offline RL needs going beyond pessimism: Representations and distributional shift. In *Decision Awareness in Reinforcement Learning Workshop at ICML 2022*.

Ghasemipour, K.; Gu, S. S.; and Nachum, O. 2022. Why so pessimistic? estimating uncertainties for offline RL through ensembles, and why their independence matters. *Advances in Neural Information Processing Systems*, 35: 18267–18281.

Goodfellow, I.; Bengio, Y.; and Courville, A. 2016. *Deep learning*. MIT press. ISBN 0262337371.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. *Advances in Neural Information Processing Systems*, 27.

Grünwald, P. D.; and Dawid, A. P. 2004. Game theory, maximum entropy, minimum discrepancy and robust Bayesian decision theory.

Gulcehre, C.; Colmenarejo, S. G.; Sygnowski, J.; Paine, T.; Zolna, K.; Chen, Y.; Hoffman, M.; Pascanu, R.; and de Freitas, N. 2020. Addressing Extrapolation Error in Deep Offline Reinforcement Learning.

Haarnoja, T.; Zhou, A.; Abbeel, P.; and Levine, S. 2018. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International Conference on Machine Learning*, 1861–1870. PMLR. ISBN 2640-3498.

Hejna, J.; Rafailov, R.; Sikchi, H.; Finn, C.; Niekum, S.; Knox, W. B.; and Sadigh, D. 2023. Contrastive prefence learning: Learning from human feedback without RL. *arXiv preprint arXiv:2310.13639*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33: 6840–6851.

Jin, Y.; Yang, Z.; and Wang, Z. 2021. Is pessimism provably efficient for offline RL? In *International Conference on Machine Learning*, 5084–5096. PMLR. ISBN 2640-3498.

Jolicoeur-Martineau, A. 2020. On relativistic f-divergences. In *International Conference on Machine Learning*, 4931–4939. PMLR. ISBN 2640-3498.

Kang, Y.; Shi, D.; Liu, J.; He, L.; and Wang, D. 2023. Beyond reward: Offline preference-guided policy optimization. *arXiv preprint arXiv:2305.16217*.

Kaufmann, T.; Weng, P.; Bengs, V.; and Hüllermeier, E. 2023. A survey of reinforcement learning from human feedback. *arXiv preprint arXiv:2312.14925*.

Kim, C.; Park, J.; Shin, J.; Lee, H.; Abbeel, P.; and Lee, K. 2023. Preference transformer: Modeling human preferences using transformers for RL. *arXiv preprint arXiv:2303.00957*.

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Knox, W. B.; Hatgis-Kessell, S.; Booth, S.; Niekum, S.; Stone, P.; and Allievi, A. 2022. Models of human preference for learning reward functions. *arXiv preprint arXiv:2206.02231*.

Kostrikov, I.; Fergus, R.; Tompson, J.; and Nachum, O. 2021. Offline reinforcement learning with Fisher divergence critic regularization. In *International Conference on Machine Learning*, 5774–5783. PMLR. ISBN 2640-3498.

Kostrikov, I.; Nair, A.; and Levine, S. 2021. Offline reinforcement learning with implicit Q-learning. *arXiv preprint arXiv:2110.06169*.

Kumar, A.; Zhou, A.; Tucker, G.; and Levine, S. 2020. Conservative Q-learning for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 33: 1179–1191.

Lange, S.; Gabel, T.; and Riedmiller, M. 2012. *Batch reinforcement learning*, 45–73. Springer.

Li, J.; Zhang, E.; Yin, M.; Bai, Q.; Wang, Y.-X.; and Wang, W. Y. 2023. Offline reinforcement learning with closed-form policy improvement operators. In *International Conference on Machine Learning*, 20485–20528. PMLR. ISBN 2640-3498.

Lu, C.; Ball, P. J.; Rudner, T. G.; Parker-Holder, J.; Osborne, M. A.; and Teh, Y. W. 2022. Challenges and opportunities in offline reinforcement learning from visual observations. *arXiv preprint arXiv:2206.04779*.

Miyato, T.; Kataoka, T.; Koyama, M.; and Yoshida, Y. 2018. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*.

Nair, A.; Gupta, A.; Dalal, M.; and Levine, S. 2020. AWAC: Accelerating online reinforcement learning with offline datasets. *arXiv preprint arXiv:2006.09359*.

Peng, X. B.; Kumar, A.; Zhang, G.; and Levine, S. 2019. Advantage-weighted regression: Simple and scalable off-policy reinforcement learning. *arXiv preprint arXiv:1910.00177*.

Peters, J.; Mulling, K.; and Altun, Y. 2010. Relative entropy policy search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 24, 1607–1612. ISBN 2374-3468.

Rafailov, R.; Sharma, A.; Mitchell, E.; Manning, C. D.; Ermon, S.; and Finn, C. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Rafailov, R.; Yu, T.; Rajeswaran, A.; and Finn, C. 2021. Offline reinforcement learning from images with latent space models. In *Learning for dynamics and control*, 1154–1168. PMLR. ISBN 2640-3498.

Razzaghi, P.; Tabrizian, A.; Guo, W.; Chen, S.; Taye, A.; Thompson, E.; Bregeon, A.; Baheri, A.; and Wei, P. 2022. A survey on reinforcement learning in aviation applications. *arXiv preprint arXiv:2211.02147*.

Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Shalev-Shwartz, S.; Shamir, O.; and Shammah, S. 2017. Failures of gradient-based deep learning. In *International Conference on Machine Learning*, 3067–3075. PMLR. ISBN 2640-3498.

Sutton, R. S.; and Barto, A. G. 2018. *Reinforcement learning: An introduction*. MIT press. ISBN 0262352702.

Swamy, G.; Dann, C.; Kidambi, R.; Wu, Z. S.; and Agarwal, A. 2024. A minimaximalist approach to reinforcement learning from human feedback. *arXiv preprint arXiv:2401.04056*.

Tarasov, D.; Kurenkov, V.; Nikulin, A.; and Kolesnikov, S. 2023. Revisiting the Minimalist Approach to Offline Reinforcement Learning. *arXiv preprint arXiv:2305.09836*.

Vieillard, N.; Andrychowicz, M.; Raichuk, A.; Pietquin, O.; and Geist, M. 2021. Implicitly regularized RL with implicit Q-values. *arXiv preprint arXiv:2108.07041*.

Wang, Z.; Hunt, J. J.; and Zhou, M. 2022. Diffusion policies as an expressive policy class for offline reinforcement learning. *arXiv preprint arXiv:2208.06193*.

Wang, Z.; Novikov, A.; Zolna, K.; Merel, J. S.; Springenberg, J. T.; Reed, S. E.; Shahriari, B.; Siegel, N.; Gulcehre, C.; and Heess, N. 2020. Critic regularized regression. *Advances in Neural Information Processing Systems*, 33: 7768–7778.

Wu, Y.; Tucker, G.; and Nachum, O. 2019. Behavior regularized offline reinforcement learning. *arXiv preprint arXiv:1911.11361*.

Yarats, D.; Fergus, R.; Lazaric, A.; and Pinto, L. 2021. Mastering visual continuous control: Improved data-augmented reinforcement learning. *arXiv preprint arXiv:2107.09645*.

Zhuang, Z.; Lei, K.; Liu, J.; Wang, D.; and Guo, Y. 2023. Behavior proximal policy optimization. *arXiv preprint arXiv:2302.11312*.

Ziebart, B. D.; Maas, A. L.; Bagnell, J. A.; and Dey, A. K. 2008. Maximum entropy inverse reinforcement learning. In *AAAI*, volume 8, 1433–1438. Chicago, IL, USA.