




Data Augmentation Through Expert-Guided Symmetry Detection to Improve Performance in Offline Reinforcement Learning

Giorgio Angelotti^{1,2}^a, Nicolas Drougard^{1,2}^b and Caroline P. C. Chanel^{1,2}^c

¹ISAE-SUPAERO, University of Toulouse, France

²ANITI, University of Toulouse, France

Keywords: Offline Reinforcement Learning, Batch Reinforcement Learning, Markov Decision Processes, Symmetry Detection, Homomorphism, Density Estimation, Data Augmenting, Normalizing Flows, Deep Neural Networks.

Abstract: Offline estimation of the dynamical model of a Markov Decision Process (MDP) is a non-trivial task that greatly depends on the data available in the learning phase. Sometimes the dynamics of the model is invariant with respect to some transformations of the current state and action. Recent works showed that an expert-guided pipeline relying on Density Estimation methods as Deep Neural Network based Normalizing Flows effectively detects this structure in deterministic environments, both categorical and continuous-valued. The acquired knowledge can be exploited to augment the original data set, leading eventually to a reduction in the distributional shift between the true and the learned model. Such data augmentation technique can be exploited as a preliminary process to be executed before adopting an Offline Reinforcement Learning architecture, increasing its performance. In this work we extend the paradigm to also tackle non-deterministic MDPs, in particular, 1) we propose a detection threshold in categorical environments based on statistical distances, and 2) we show that the former results lead to a performance improvement when solving the learned MDP and then applying the optimized policy in the real environment.


1 INTRODUCTION


In Offline Reinforcement Learning (ORL) and Offline Learning for Planning the environment dynamics and/or value functions are inferred from a batch of already pre-collected experiences. Wrong previsions lead to bad decisions. The distributional shift, defined as the discrepancy between the learnt model and reality, is the main responsible for the performance deficit of the (sub)optimal policy obtained in the offline setting compared to the true optimal policy (Levine et al., 2020; Angelotti et al., 2020). Is there a way to exploit expert knowledge or intuition about the environment to limit the distributional shift? Several models benefit from a dynamics that is invariant with respect to some transformations of the system of reference. In physics, such a property of a system is called a symmetry (Gross, 1996). In the context of Markov Decision Processes (MDPs) (Bellman, 1966) a symmetry can be defined as a particular case of an MDP's homomor-


phism (Angelotti et al., 2022). Knowing that a system to be learned is endowed with a symmetry or of a homomorphic structure can lead to more data-efficient solutions of an MDP.

The automatic discovery of homomorphic structures in MDPs has a long story (Dean and Givan, 1997; Ravindran and Barto, 2001; Ravindran and Barto, 2004). In (Li et al., 2006) a theoretical analysis of the possible types of MDPs state abstractions proved which properties of the original MDP would be invariant under the transformation: the optimal value function, the optimal policy, etc. Eventually, the full automatic discovery of a factored MDP representation was proven to be as hard as verifying whether two graphs are isomorphic (Narayanamurthy and Ravindran, 2008). In recent years (van der Pol et al., 2020a; van der Pol et al., 2020b; Angelotti et al., 2022) rekindled the topic.

In (van der Pol et al., 2020a) a contrastive loss function that enforces action equivariance on a to-be-learned representation of an MDP was adopted to learn a structured latent space that was then exploited to increase the data efficiency of a data-driven planner. (van der Pol et al., 2020b) introduced peculiar

^a <https://orcid.org/0000-0002-1878-5833>

^b <https://orcid.org/0000-0003-0002-9973>

^c <https://orcid.org/0000-0003-3578-4186>

classes of Deep Neural Network (DNN) architectures that by construction enforce the invariance of the optimal MDP policy under some set of transformations obtained through other Deep RL paradigms. The latter also provided an increase in data efficiency. In (Angelotti et al., 2022) an expert-guided detection of alleged symmetries based on Density Estimation statistical techniques in the context of the offline learning of both continuous and categorical environments was proposed in order to eventually augment the starting data set. The authors showed that correctly detecting a symmetry (based on the computation of a symmetry confidence value $v_k > v$) and data augmenting the starting data set exploiting this information led to a decrease in the distributional shift. Unfortunately, the said work concerned only *deterministic* MDPs and did not include an analysis of the *performance* of the policy obtained in the end. In other fields of Machine Learning data augmentation has been extensively exploited to boost the efficiency of the algorithms in data-limited setups (van Dyk and Meng, 2001; Shorten and Khoshgoftaar, 2019; Park et al., 2019).

Recently (Yarats et al., 2022) showed the importance of large and diverse datasets for ORL by demonstrating empirically that offline learning using a vanilla online RL algorithm over a batch that is diverse enough can lead to performances that are comparable to, or even better than, pure ORL approaches.

In this context, the present work addresses the following research questions: *Is it possible to develop a method for expert-guided detection of alleged symmetries based on Density Estimation statistical techniques in the context of offline learning that also works for stochastic MDPs?* The main idea is to extend previous works (van der Pol et al., 2020a; van der Pol et al., 2020b; Angelotti et al., 2022) to deal with stochastic MDPs; and, *Is Data Augmentation exploiting a detected symmetry really beneficial to the learning of an MDP policy in the offline context?* We would like to empirically demonstrate (O)RL policy improvement when enriching the batch as proposed by (Yarats et al., 2022).

Contributions. In this work, we take over and extend the state-of-the-art with the aim of providing an answer to the listed research questions. More specifically, the contributions of this paper are the followings:

1. *Algorithmic Contribution.* A refinement of the decision threshold, based on statistical distances, is defined for categorical MDPs. This new decision threshold is valid also in both stochastic and deterministic environments, improving hence over the state-of-the-art that only tackled deterministic scenarios;
2. *Experimental Contribution.* The improvement of the policy performance obtained by augmenting the data with the symmetric images of the transitions is demonstrated experimentally in an offline learning context. The good quality of the method is clear in the categorical setting while it is fuzzier in the continuous setting since offline methods with Deep Neural Networks are affected by the (non-trivial) choice of the hyperparameters.

It is worth saying that the presented work aim is not to be a competitor to the ORL algorithms, but a way to augment the batch by validating expert intuition. Once the batch has been augmented one could use any offline RL method.

2 BACKGROUND

Definition 1 (Markov Decision Process). An MDP (Bellman, 1966) is a tuple $\mathcal{M} = (S, A, R, T, \gamma)$. S and A are the sets of states and actions, $R : S \times A \rightarrow \mathbb{R}$ is the reward function, $T : S \times A \rightarrow \text{Dist}(S)$ is the transition function, where $\text{Dist}(S)$ is the set of probability distributions on S , and $\gamma \in [0, 1)$ is the discount factor. Time is discretized and at each step $t \in \mathbb{N}$ the agent observes a system state $s = s_t \in S$, acts with $a = a_t \in A$ drawn from a policy $\pi : S \rightarrow \text{Dist}(A)$, and with probability $T(s, a, s')$ transits to a next state $s' = s_{t+1}$, earning a reward $R(s, a)$. The value function of π and s is defined as the expected total discounted reward using π and starting with s : $V_\pi(s) = \mathbb{E}_\pi[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t) | s_0 = s]$. The optimal value function V^* is the maximum of the latter over every policy π .

Definition 2 (MDP Symmetry). Given an MDP \mathcal{M} , let k be a surjection on $S \times A \times S$ such that $k(s, a, s') = (k_\sigma(s, a, s'), k_\alpha(s, a, s'), k_{\sigma'}(s, a, s')) \in S \times A \times S$. Let $(T \circ k)(s, a, s') = T(k(s, a, s'))$. k is a symmetry if $\forall (s, s') \in S^2, a \in A$ both T and R are invariant with respect to the image of k :

$$(T \circ k)(s, a, s') = T(s, a, s'), \quad (1)$$

$$R(k_\sigma(s, a, s'), k_\alpha(s, a, s')) = R(s, a). \quad (2)$$

As (Angelotti et al., 2022), in this paper we will focus only on the invariance of T , therefore we will only demand for the validity of Equation 1. Problems with a known reward function as well as model-based approaches can thus benefit directly from the method.

Probability Mass Function Estimation for Discrete MDPs. Let $\mathcal{D} = \{(s_i, a_i, s'_i)\}_{i=1}^n$ be a batch of recorded transitions. Performing mass estimation over \mathcal{D} amounts to compute the probabilities that define the

categorical distribution T by estimating the frequencies of transition in \mathcal{D} . In other words:

$$\hat{T}(s, a, s') = \begin{cases} \frac{n_{s,a,s'}}{\sum_{s'} n_{s,a,s'}} & \text{if } \sum_{s'} n_{s,a,s'} > 0, \\ |S|^{-1} & \text{otherwise.} \end{cases} \quad (3)$$

where $n_{s,a,s'}$ is the number of times the transition ($s_t = s, a_t = a, s_{t+1} = s'$) appears in \mathcal{D} .

Probability Density Function Estimation for Continuous MDPs. Performing density estimation over \mathcal{D} means obtaining an analytical expression for the probability density function (pdf) of transitions (s, a, s') given \mathcal{D} : $\mathcal{L}(s, a, s' | \mathcal{D})$. Normalizing flows (Dinh et al., 2015; Kobyzev et al., 2020) allow defining a parametric flow of continuous transformations that reshapes a known initial pdf to one that best fits the data.

Expert-Guided Detection of Symmetries. The paradigm described in (Angelotti et al., 2022) can be resumed as follows:

1. An expert presumes that a to be learned model is endowed with the invariance of T with respect to a transformation k ;
2. She/He computes the probability function estimation based on the batch \mathcal{D} :
 - (a) (categorical case) She/He computes \hat{T} , an estimate of T , using the transitions in a batch \mathcal{D} by applying Equation 3;
 - (b) (continuous case) She/He performs Density Estimation over \mathcal{D} using Normalizing Flows;
3. She/He applies k to all transitions $(s, a, s') \in \mathcal{D}$ and then checks whether the symmetry confidence value v_k ;
 - (a) (categorical case) of samples $k(s, a, s') = (k_\sigma(s, a, s'), k_\alpha(s, a, s'), k_{\sigma'}(s, a, s')) \in k(\mathcal{D})$ s.t. $T(s, a, s') = (T \circ k)(s, a, s')$ exceeds an expert given threshold v ;
 - (b) (continuous case) of probability values \mathcal{L} evaluated on $k(\mathcal{D})$ exceeds a threshold θ that corresponds to the q -order quantile of the distribution of probability values evaluated on the original batch. The quantile order q is given as an input to the procedure by an expert (see Algorithm 2);
4. If the last condition is fulfilled then \mathcal{D} is augmented with $k(\mathcal{D})$.

Note that once a transformation k is detected as a symmetry the dataset is potentially augmented with transitions that are not present in the original batch, injecting hence unseen and totally novel information into the dataset.

3 ALGORITHMIC CONTRIBUTION

Our algorithmic contribution consists in the improvement of the calculation of v_k in part (3.a) of the previous list (Angelotti et al., 2022). Indeed, that approach does not yield valid results when applied to stochastic environments. In order for the method to work in stochastic environments we need to measure a distance in distribution. The latter somehow was considered in the version of the approach that took care of continuous deterministic environments since learning a distribution over transitions represented by their features is independent of the nature of the dynamics. However, when dealing with categorical states the notion of distance between features can't be exploited.

We propose to compute the percentage v_k relying on a distance between categorical distributions. Since the transformation k is a surjection on transition tuples, we do not know a-priori which will be the correct mapping $k_{\sigma'}(s, a, s') \forall s' \in S$. In other words, we can compute $k_{\sigma'}$, the symmetric image of s' , only when we receive as an input the whole tuple (s, a, s') since an inverse mapping might not exist.

Therefore we will resort to computing a *pessimistic* approximation of the Total Variational Distance (proportional to the L^1 -norm). In particular, given (s, a, s') , we aim to calculate the Chebyshev distance (the L^∞ -norm) between $T(s, a, \cdot)$ and $T(k_\sigma(s, a, s'), k_\alpha(s, a, s'), \cdot)$. Recall that given two vectors of dimension d , x and y both $\in \mathbb{R}^d$, $\|x - y\|_\infty \leq \|x - y\|_1$.

Let us then define the following four functions:

$$m(s, a, s') = \min_{\bar{s} \in S \setminus \{s'\}: \hat{T} \neq 0} \hat{T}(s, a, \bar{s}) \quad (4)$$

$$M(s, a, s') = \max_{\bar{s} \in S \setminus \{s'\}} \hat{T}(s, a, \bar{s}), \quad (5)$$

$$m_k(s, a, s') = \min_{\substack{\bar{s} \in S \text{ s.t.} \\ \bar{s} \neq k_{\sigma'}(s, a, s') \\ \text{and } \hat{T} \circ k \neq 0}} \hat{T}(k_\sigma(s, a, s'), k_\alpha(s, a, s'), \bar{s}), \quad (6)$$

$$M_k(s, a, s') = \max_{\substack{\bar{s} \in S \text{ s.t.} \\ \bar{s} \neq k_{\sigma'}(s, a, s')}} \hat{T}(k_\sigma(s, a, s'), k_\alpha(s, a, s'), \bar{s}) \quad (7)$$

where m (M) and m_k (M_k) are the minimum (maximum) of the probability mass function (pmf) \hat{T} when evaluated respectively on an initial state and action (s, a) and $(k_\sigma(s, a, s'), k_\alpha(s, a, s'))$ for which $\hat{T} \neq 0$. Those zero values are excluded because, in the context of a small dataset, many transitions are unexplored, and including values = 0 would often lead to over-pessimistic estimates.

In order to approximate the Chebyshev distance between $\hat{T}(s, a, \cdot)$ and $\hat{T}(k_\sigma(s, a, s'), k_\alpha(s, a, s'), \cdot)$ we

define a pessimistic approximation d_k as follows:

$$d_k(s, a, s') = \max \left\{ \underbrace{|M(s, a, s') - m_k(s, a, s')|}_{(I)}, \right. \\ \left. \underbrace{|M_k(s, a, s') - m(s, a, s')|}_{(II)}, \right. \\ \left. \underbrace{|\hat{T}(s, a, s') - (\hat{T} \circ k)(s, a, s')|}_{(III)} \right\}. \quad (8)$$

For the moment consider $\hat{T}(s, a, \cdot)$ and $\hat{T}(k_\sigma(s, a, s'), k_\alpha(s, a, s'), \cdot)$ just as two sets of numbers. Remove the value corresponding to s' from the first set, the one corresponding to $k_{\sigma'}(s, a, s')$ from the second set, and any remaining zeros from both. Taking the max between (I) and (II) just equates to selecting the maximum possible difference between any two values of these modified sets. Equation 8 simply tells us to select the worst possible case since we do not know which permutations of states we should compare when computing the Chebyshev distance. s' is removed from $\hat{T}(s, a, \cdot)$ and $k_{\sigma'}(s, a, s')$ is removed from $\hat{T}(k_\sigma(s, a, s'), k_\alpha(s, a, s'), \cdot)$ since we know that k maps (s, a, s') to $(k_\sigma(s, a, s'), k_\alpha(s, a, s'), k_{\sigma'}(s, a, s'))$ and hence we can compare those values directly (III).

Notice that

$$0 < d_k(s, a, s') \leq 1 \quad \forall (s, a, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}. \quad (9)$$

In the following, we propose to improve the algorithms proposed in (Angelotti et al., 2022). In detail, we redefine the symmetry confidence value v_k . We propose to estimate v_k as in Line 2 of Algorithm 1 as:

$$v_k(\mathcal{D}) = 1 - \frac{1}{|\mathcal{D}|} \sum_{(s, a, s') \in \mathcal{D}} d_k(s, a, s'). \quad (10)$$

From equations 8 and 9, it follows that: (i) in deterministic environments v_k (Eq. 10) coincides with the one prescribed in (Angelotti et al., 2022); and, (ii) $1 > v_k \geq 0$, so v_k can be interpreted as a percentage. This last allows us to suppose that v_k is an estimate of the probability of k being a symmetry of the dynamics, and therefore we can relax the necessity of defining an expert-given threshold v (cf. (Angelotti et al., 2022) Alg. 1). We then set $v = 0.5$ as an input in Algorithm 1 and eventually augment the batch if $v_k > 0.5$ (Lines 3-5).

Remark (Extreme Case Scenario). Is Equation 8 too pessimistic? Consider that for a given state action couple (\bar{s}, \bar{a}) we have a transition distributed over 3 states $s \in \mathcal{S} = \{One, Two, Three\}$ with probabilities $T(\bar{s}, \bar{a}, One) = 0.01$, $T(\bar{s}, \bar{a}, Two) = 0.01$ and $T(\bar{s}, \bar{a}, Three) = 0.98$. Now, assume the estimate of

Algorithm 1: Symmetry detection and data augmenting in a categorical MDP.

Input: Batch of transitions \mathcal{D} , k alleged symmetry
Output: Possibly augmented batch $\mathcal{D} \cup \mathcal{D}_k$

- 1 $\hat{T} \leftarrow$ Most Likely Categorical pmf from \mathcal{D}
- 2 $v_k = 1 - \frac{1}{|\mathcal{D}|} \sum_{(s, a, s') \in \mathcal{D}} d_k(s, a, s')$ (where d_k is defined in Equation 8)
- 3 **if** $v_k > 0.5$ **then**
- 4 $\mathcal{D}_k = k(\mathcal{D})$ (alleged symm. transitions)
- 5 **return** $\mathcal{D} \cup \mathcal{D}_k$ (the augmented batch)
- 6 **else**
- 7 **return** \mathcal{D} (the original batch)
- 8 **end**

the transition function is perfect. Does the distance in Equation 8 converge to 0? Not always, but what matters for the detection of symmetries is the average of the distances over the whole batch (Eq. 10). Suppose that these probabilities were inferred from a batch with the transition (\bar{s}, \bar{a}, One) once, (\bar{s}, \bar{a}, Two) once and $(\bar{s}, \bar{a}, Three)$ ninety-eight times. Consider $(\bar{s}, \bar{a}, Three)$. $M(\bar{s}, \bar{a}, Three) = M_k(\bar{s}, \bar{a}, Three) = m(\bar{s}, \bar{a}, Three) = m_k(\bar{s}, \bar{a}, Three) = 0.01$. Following Eq. 8, $d_k(\bar{s}, \bar{a}, Three) = 0$. However, $d_k(\bar{s}, \bar{a}, One) = d_k(\bar{s}, \bar{a}, Two) = 0.97$, which is a too pessimistic estimate. Nevertheless let's calculate v_k (Eq.10). For this state-action pair (\bar{s}, \bar{a}) , the average over the batch is therefore: $(d_k(\bar{s}, \bar{a}, One) + d_k(\bar{s}, \bar{a}, Two) + 98d_k(\bar{s}, \bar{a}, Three))/100 = 0.0194$. If the estimation is the same for other pairs (s, a) , then $v_k = 1 - 0.0194 = 0.9806$. This is a value close to 1 suggesting k is a symmetry.

4 EXPERIMENTS

In order to show the improvements provided by our contribution we tested the algorithms in a stochastic version of the toroidal Grid environment and two continuous state environments of the OpenAI's Gym Learning Suite: CartPole and Acrobot. We have chosen the same scenarios as (Angelotti et al., 2022) in order to demonstrate that our approach generalizes well to the stochastic case contrary to the approach proposed in (Angelotti et al., 2022).

4.1 Setup

We collect a batch of transitions \mathcal{D} using a uniform random policy. An expert alleges the presence of a symmetry k and we proceed to its detection using

Algorithm 2: Symmetry detection and data augmenting in a continuous MDP with detection threshold $v = 0.5$ (Angelotti et al., 2022).

Input: Batch of transitions \mathcal{D} , $q \in [0, 1)$ order of the quantile, k alleged symmetry
Output: Possibly augmented batch $\mathcal{D} \cup \mathcal{D}_k$

- 1 $\mathcal{L} \leftarrow$ Density Estimate (\mathcal{D}) (e.g. with *Normalizing Flows*)
- 2 $\Lambda \leftarrow$ Distribution $\mathcal{L}(\mathcal{D})$ (\mathcal{L} evaluated over \mathcal{D})
- 3 $\theta = q$ -order quantile of Λ
- 4 $\mathcal{D}_k = k(\mathcal{D})$ (alleged symmetric transitions)
- 5 $v_k = \frac{1}{|\mathcal{D}_k|} \sum_{(s,a,s') \in \mathcal{D}_k} \mathbb{1}_{\{\mathcal{L}(s,a,s'|\mathcal{D}) > \theta\}}$
- 6 **if** $v_k > 0.5$ **then**
- 7 | **return** $\mathcal{D} \cup \mathcal{D}_k$ (*the augmented batch*)
- 8 **else**
- 9 | **return** \mathcal{D} (*the original batch*)
- 10 **end**

Algorithm 1 (categorical case) or Algorithm 2 (continuous case). In the continuous case, Density Estimation is performed by a Masked Autoregressive Flow architecture (Papamakarios et al., 2017) with 3 layers of bijectors.

The experiments were performed using 2 Dodeca-core Skylake Intel® Xeon® Gold 6126 @ 2.6 GHz and 96 GB of RAM and 2 GPU NVIDIA® V100 @ 192GB of RAM. The code to run the experiments is available at <https://github.com/giorgioangel/dsym>.

Computation of v_k and Batch Augmentation. We report the v_k obtained with an ensemble of N different iterations of the procedure: we generate $z \in \mathbb{N}$ sets of N different batches \mathcal{D} of increasing size. Remember that since $v_k \in [0, 1)$ we can interpret it as the probability of the presence of a symmetry and select a detection threshold $v = 0.5$ or higher, while in (Angelotti et al., 2022) the threshold v was expert-given. We calculate v_k with both the (Angelotti et al., 2022) method and the approach here presented.

Evaluation of the Performance (Categorical Case).

In the end, let ρ be the distribution of initial states $s_0 \in S$ and let the performance U^π of a policy π be $U^\pi = \mathbb{E}_{s \sim \rho} [V^\pi(s)]$. Our experimental contribution is the comparison between the performances obtained by acting in the real environment with $\hat{\pi}$ (the optimal policy solving the MDP defined with \hat{T}) and $\hat{\pi}_k$ (the optimal policy obtained with \hat{T}_k). In particular we consider the quantity

$$\Delta U = U^{\hat{\pi}_k} - U^{\hat{\pi}}. \quad (11)$$

$\Delta U > 0$ means that data augmenting leads to better policies.

In *categorical* environments the policies are obtained with Policy Iteration and evaluated with Policy Evaluation.

Evaluation of the Performance (Continuous Case).

In *continuous* environments Offline Learning is not trivial. We use the implementation of two Model-Free Deep RL architectures: Deep Q-Network (DQN) (Mnih et al., 2015) and Conservative Q-Learning (CQL) (Kumar et al., 2020) of the d3rlpy learning suite (Seno and Imai, 2021) to obtain a policy starting from the batches. The first method is the one that originally established the validity of Deep RL and it is used in online RL while the second was specifically developed to tackle offline RL problems. Since the convergence of the training of Deep RL baselines is greatly dependent of hyperparameter tuning that itself depends on both the environment and the batch (Paine et al., 2020), we will apply DQN and CQL with the default parameters provided by d3rlpy, abiding hence more faithfully to an offline learning duty. This means that sometimes the learning might not converge to a good policy. We find this philosophy more honest than showing the results obtained with the best seed or the finest-tuned hyperparameters. Each architecture is trained for a number of steps equal to fifty times the number of transitions present in the batch.

4.2 Environments

Stochastic Grid (Categorical). In this environment, the agent can move along fixed directions over a torus by acting with any $a \in A = \{\uparrow, \downarrow, \leftarrow, \rightarrow\}$ (see Figure 1). The grid meshing the torus has size $l = 10$.

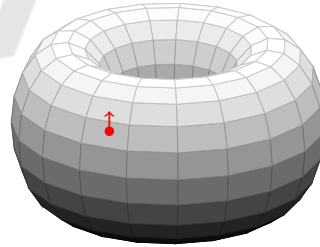


Figure 1: Representation of the Grid Environment (Angelotti et al., 2022). The red dot is the position of a state s on the torus. A possible displacement obtained by acting with action $a = \uparrow$ is shown as a red arrow.

The agent can spawn everywhere on the torus with a uniform probability and must reach a fixed goal. At every time step, the agent receives a reward $r = -1$ if it does not reach the goal and a reward $r = 1$ once the goal has been reached, terminating the episode. When performing an action the agent has 60% chances of moving to the intended direction, 20% to the opposite

one, and 10% along an orthogonal direction. We collect $z = 10$ sets of $M = 100$ batches with respectively $N = 1000 \times i_z$ steps in each batch (i_z going from 1 to z).

Table 1: Toroidal Grid: proposed transformations and label.

k	Label
$k_G(s, a, s') = s'$	TRSAI
$k_A(s, a = (\uparrow, \downarrow, \leftarrow, \rightarrow), s') = (\downarrow, \uparrow, \rightarrow, \leftarrow)$	
$k_G(s, a, s') = s$	
$k_G(s, a, s') = s$	SDAI
$k_A(s, a = (\uparrow, \downarrow, \leftarrow, \rightarrow), s') = (\downarrow, \uparrow, \rightarrow, \leftarrow)$	
$k_G(s, a, s') = s'$	
$k_G(s, a, s') = s$	ODAI
$k_A(s, a = (\uparrow, \downarrow, \leftarrow, \rightarrow), s') = (\downarrow, \uparrow, \rightarrow, \leftarrow)$	
$k_G(s, a = (\uparrow, \downarrow, \leftarrow, \rightarrow), s') = (s' - (0, 2), s' + (0, 2), s' + (2, 0), s' - (2, 0))$	
$k_G(s, a, s') = s$	ODWA
$k_A(s, a = (\uparrow, \downarrow, \leftarrow, \rightarrow), s') = (\rightarrow, \leftarrow, \uparrow, \downarrow)$	
$k_G(s, a = (\uparrow, \downarrow, \leftarrow, \rightarrow), s') = (s' - (0, 2), s' + (0, 2), s' + (2, 0), s' - (2, 0))$	
$k_G(s, a, s') = s'$	TI
$k_A(s, a, s') = a$	
$k_G(s, a = (\uparrow, \downarrow, \leftarrow, \rightarrow), s') = (s' + (0, 1), s' - (0, 1), s' - (1, 0), s' + (1, 0))$	
$k_G(s, a, s') = s'$	TIOD
$k_A(s, a, s') = a$	
$k_G(s, a, s') = s$	

The proposed symmetries for this environment are outlined in Table 1. We check for the invariant of the dynamics with respect to the following six transformations (the valid symmetries are displayed in **bold**): (1) Time reversal symmetry with action inversion (**TRSAI**); (2) Same dynamics with action inversion (SDAI); (3) Opposite dynamics and action inversion (**ODAI**); (4) Opposite dynamics but wrong action (ODWA); (5) Translation invariance (**TI**); (6) Translation invariance with opposite dynamics (TIOD). The N dependent average results for symmetry detection using the method from (Angelotti et al., 2022) are reported in Figure 2, and results using our method are displayed in Figure 3a. Figure 3b presents the performance improvement ΔU , with its standard deviation being represented by a vertical error bar.

Stochastic CartPole (Continuous). A pole is precariously balanced on a cart and an agent can push the whole system left or right to prevent the pole from falling.

The dynamics is similar to that of CartPole (Brockman et al., 2016) (see Figure 5), however the force that the agent uses to push the cart is sampled from a normal distribution with mean f (the force defined in the deterministic version) and standard deviation $\tilde{\sigma} = 2$. Recall that the state is represented by the features (x, θ, v, ω) and $A = \{\leftarrow, \rightarrow\}$. For the evaluation of v_k

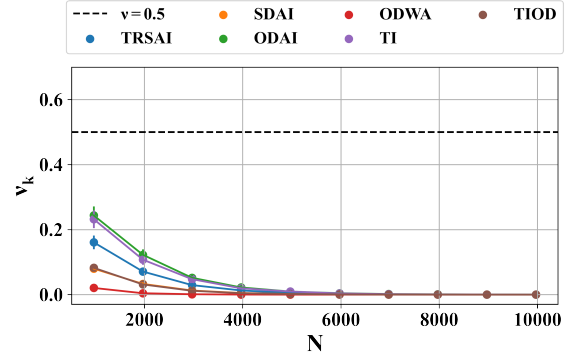
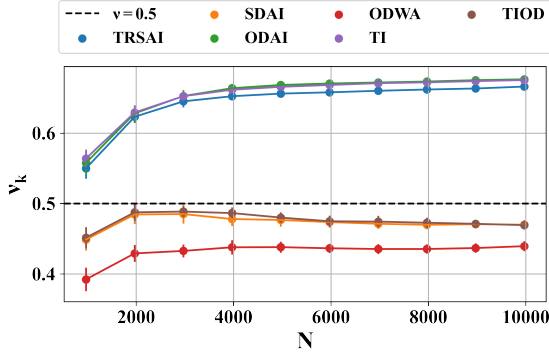


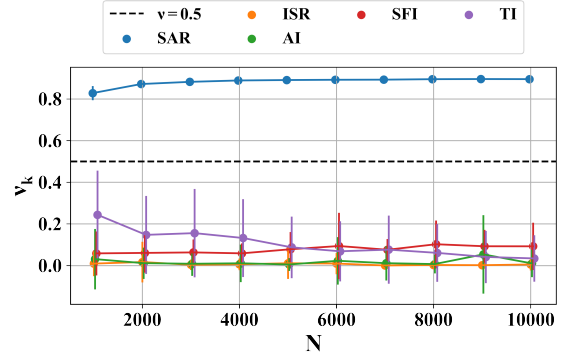
Figure 2: Stochastic toroidal Grid Environment. Probability of symmetry v_k calculated with the method proposed by (Angelotti et al., 2022). The threshold at $v = 0.5$ is displayed as a dashed line. Since all $v_k < 0.5$ means that no transformation is detected as a symmetry.

we set the quantile $q = 0.1$ and we collect $z = 10$ sets of $M = 100$ batches with respectively $N = 1000 \times i_z$ steps in each batch (and i_z going from 1 to 10). We evaluate ΔU by training the agent on single batches of $N = 5000 \times i_z$ (and i_z going from 1 to 6) both augmented and not augmented with k . The acronyms of the valid symmetric transformations are displayed in **bold**: (1) State and action reflection with respect to an axis in $x = 0$ (**SAR**); (2) Initial state reflection (ISR); (3) Action inversion (AI); (4) Single feature inversion (SFI); (5) Translation invariance (**TI**). Their effects on the transition (s, a, s') are listed in Table 2. Average results and errors are displayed in Figure 4a. The results considering the evaluation of performance gain (ΔU) are shown in Table 4.

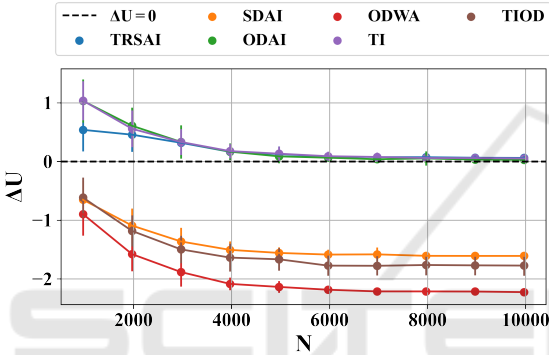
Stochastic Acrobot (Continuous). The Acrobot is a planar two-link robotic arm working against gravity, the agent can decide whether to swing or not the elbow left or right to balance the arm straightened up (see Figure 6). It is the very same Acrobot of (Brockman et al., 2016) but at every time step a noise ε is sampled from a uniform distribution on the interval $[-0.5, 0.5]$ and added to the torque. A state is represented by the features $(s_1, c_1, s_2, c_2, \omega_1, \omega_2)$ where s_i and c_i are respectively $\sin(\alpha_i)$ and $\cos(\alpha_i)$ in shorthand notation. The action set $A = \{-1, 0, 1\}$. For the evaluation of v_k we set $q = 0.1$. For the detection case, we collected $z = 5$ sets of $M = 100$ batches with $N = 1000 \times i_z$ steps within each one (i_z going from 1 to z). The evaluation of the performance was carried out on single batches, with and without data augmentation, with $N = 10000 \times i_z$ steps and i_z going from 1 to 4. For the evaluation of $\Delta z = 5$ due to computational necessities. We allege the following transformations k , as always the valid ones are **bolded**: (1) Angles and angular



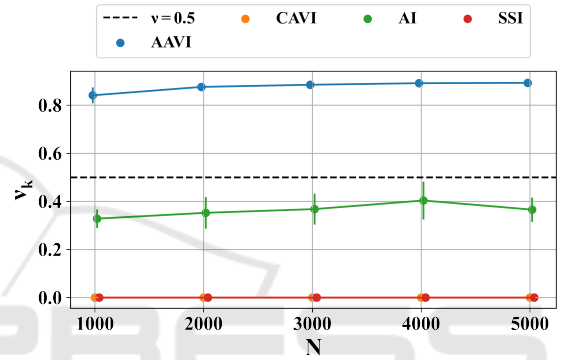
(a) Probability of symmetry v_k **with our approach**. The threshold at $v = 0.5$ is displayed as a dashed line. $v_k > 0.5$ means that the transformation is detected as a symmetry.



(a) Stochastic CartPole. Probability of symmetry v_k . The threshold at $v = 0.5$ is displayed as a dashed line. $v_k > 0.5$ means that the transformation is detected as a symmetry.



(b) Performance difference ΔU (Eq. 11). The threshold at $\Delta U = 0$ is displayed as a dashed line. $\Delta U > 0$ means that data augmenting leads to better policies.



(b) Stochastic Acrobot. Probability of symmetry v_k . The threshold at $v = 0.5$ is displayed as a dashed line. $v_k > 0.5$ means that the transformation is detected as a symmetry.

Figure 3: Stochastic Toroidal Grid Environment. v_k and ΔU for the transformations k computed over sets of 100 different batches of size N . Points are mean values and bars standard deviations.

velocities inversion (**AAVI**); (2) Cosines and angular velocities inversion (**CAVI**); (3) Action inversion (**AI**); (4) Starting state inversion (**SSI**).

The images of the transformations are reported in Table 3. The N dependent average results and standard deviations are reported in Figure 4b. The results considering the evaluation of performance gain (ΔU) are shown in Table 5.

5 DISCUSSION

Stochastic Grid (Categorical). *Detection phase* (v_k). We see from Figure 2 that using the state-of-the-art approach no transformation is detected as a symmetry because $v_k < 0.5, \forall k$ in the proposed set of transformations. This result highlights the inadequacy

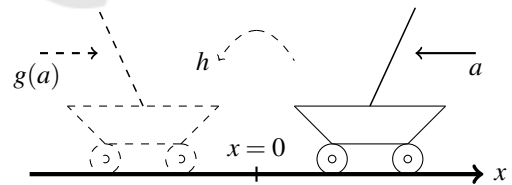


Figure 5: The cart in the right is a representation of a Cart-Pole’s state s_t with $x_t > 0$ and action $a_t \leftarrow$ (Angelotti et al., 2022). The dashed cart in the left is the image of (s_t, a_t) under the transformation h which inverts state $f(s) = -s$ and action $g(a) = -a$.

of the state-of-the-art method to deal with stochastic environments. On the contrary, our novel algorithm perfectly manages to identify the real symmetries of the environment (see Figure 3a): $v_k > 0.5, k \in \{\text{TRSAI}, \text{ODAI}, \text{TI}\}$. Moreover, there are no false positives: $v_k < 0.5, k \in \{\text{SDAI}, \text{ODWA}, \text{TIOD}\}$. We

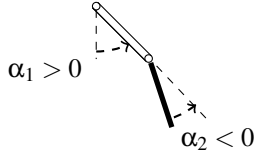


Figure 6: Representation of a state of the Acrobot environment (Angelotti et al., 2022).

Table 2: Proposed transformations and labels for Stochastic CartPole.

k	Label
$k_{\sigma}(s, a, s') = -s$ $k_{\alpha}(s, a = (\leftarrow, \rightarrow), s') = (\rightarrow, \leftarrow)$ $k_{\sigma'}(s, a, s') = -s'$	SAR
$k_{\sigma}(s, a, s') = -s$ $k_{\alpha}(s, a, s') = a$ $k_{\sigma'}(s, a, s') = s'$	ISR
$k_{\sigma}(s, a, s') = s$ $k_{\alpha}(s, a = (\leftarrow, \rightarrow), s') = (\rightarrow, \leftarrow)$ $k_{\sigma'}(s, a, s') = s'$	AI
$k_{\sigma}(s = (x, \dots), a, s') = (-x, \dots)$ $k_{\alpha}(s, a, s') = a$ $k_{\sigma'}(s, a, s') = s'$	SFI
$k_{\sigma}(s = (x, \dots), a, s') = (x + 0.3, \dots)$ $k_{\alpha}(s, a, s') = a$ $k_{\sigma'}(s, a, s') = (s' + 0.3, \dots)$	TI

notice that while in a deterministic environment $v_k = 0 \forall k$ which is not a symmetry, here the stochasticity makes the detection more complicated since $v_k \approx 0.5$ for $N = 2000$.

Evaluation of performance gain (ΔU). The difference in the performance of the deployed policies

Table 3: Proposed transformations and labels for Stochastic Acrobot.

k	Label
$k_{\sigma}(s = (s_1, s_2, \omega_1, \omega_2, \dots), a, s')$ $= (-s_1, -s_2, -\omega_1, -\omega_2, \dots)$ $k_{\alpha}(s, a = (-1, 0, 1), s') = (1, 0, -1)$ $k_{\sigma'}(s, a, s' = (s'_1, s'_2, \omega'_1, \omega'_2, \dots))$ $= (-s'_1, -s'_2, -\omega'_1, -\omega'_2, \dots)$	AAVI
$k_{\sigma}(s = (c_1, c_2, \omega_1, \omega_2, \dots), a, s')$ $= (-c_1, -c_2, -\omega_1, -\omega_2, \dots)$ $k_{\alpha}(s, a = (-1, 0, 1), s') = (1, 0, -1)$ $k_{\sigma'}(s, a, s' = (c'_1, c'_2, \omega'_1, \omega'_2, \dots))$ $= (-c'_1, -c'_2, -\omega'_1, -\omega'_2, \dots)$	CAVI
$k_{\sigma}(s, a, s') = s$ $k_{\alpha}(s, a = (-1, 0, 1), s') = (1, 0, -1)$ $k_{\sigma'}(s, a, s') = s'$	AI
$k_{\sigma}(s, a, s') = -s$ $k_{\alpha}(s, a, s') = a$ $k_{\sigma'}(s, a, s') = s'$	SSI

ΔU perfectly fits the expected behavior. When k is a symmetry $\Delta U > 0$ and saturates to 0 with N increasing. When k is not a symmetric transformation of the dynamics $\Delta U < 0$ and keeps decreasing with N (see Figure 3b).

Stochastic CartPole (Continuous). *Detection phase (v_k)* In Stochastic CartPole the algorithm fails to detect the symmetry $k = \text{TI}$. This could be due to the fact that the translation invariance symmetry in this case is fixed for a specific value (see TI in Table 2 where the translation is set at 0.3). If the translation is too small the neural network fails to discern the transformation from the noise. The algorithm classifies correctly as a symmetry $k = \text{SAR}$ and the remaining transformations as non-symmetries (see Figure 4a).

Evaluation of performance gain (ΔU). Results are displayed in Table 4. ORL is very unstable and sensitive to the choice of hyperparameters. On top of that, the training is carried out for a fixed number of epochs. We notice that, on average over different batch sizes, $\Delta U > 0$ for DQN and SAR, and SFI transformations. While SAR is a valid symmetry, SFI it's not. A more conservative algorithm like CQL only detects SAR as a valid symmetry. The performance difference for TI both for DQN and CQL is so close to zero that we think that augmenting the dataset with this symmetry might not be a substantial power-up over using just the information contained in the original batch.

Stochastic Acrobot (Continuous). *Detection phase (v_k).* In this environment the only real symmetry of the dynamics, **AAVI**, gets successfully detected by the algorithm with $q = 0.1$. Non symmetries yield a $v_k < 0.5$ (Figure 4b).

Evaluation of performance gain (ΔU). Results are displayed in Table 5 and show that the training in Stochastic Acrobot is harder than in Stochastic CartPole since, even with a large dataset, sometimes the algorithms do not manage to learn a good policy. In particular, while CQL manages to learn how to behave in the environment exploiting the **AAVI** symmetry (average $\Delta U = 52.9$), DQN still struggles with every k , good and wrong. Nevertheless, CQL apparently benefits from augmenting the dataset also with wrong symmetries even though to a smaller extent. We suppose this effect is due to the instability in ORL training.

6 CONCLUSIONS

Data efficiency in the offline learning of MDPs is highly coveted. Exploiting the intuition of an expert

Table 4: ΔU for every alleged symmetry in Stochastic CartPole with two baselines and different batch sizes N .

k	Baseline	N (number of transitions in the original batch)						Average ΔU
		5000	10000	15000	20000	25000	30000	
SAR	DQN	-7.3	25.4	41.8	7.2	9.0	3.4	13.3
	CQL	37.4	-2.5	-4.1	20.1	17.9	-9.0	10.0
ISR	DQN	-1.3	-48.5	-29.9	-78.7	-107.8	-29.1	-49.2
	CQL	6.4	1.6	-2.2	-22.3	-10.3	-25.9	-8.8
AI	DQN	26.9	-48.5	-43.7	-74.6	-41.3	-84.6	-44.3
	CQL	-13.1	-7.6	-29.8	-6.5	-22.3	-15.3	-15.8
SFI	DQN	-33.4	17.9	21.4	45.4	-6.9	-0.1	7.4
	CQL	-5.5	-2.1	7.4	-3.9	-3.6	-18.5	-4.4
TI	DQN	36.9	-28.1	34.5	15.7	6.1	-9.1	-0.2
	CQL	7.6	-1.3	-2.1	11.8	-16.5	5.2	0.8

Table 5: ΔU for every alleged symmetry in Stochastic Acrobot with two baselines and different batch sizes N .

k	Baseline	N				Average ΔU
		10000	20000	30000	40000	
AAVI	DQN	24.7	-17.5	-63.4	-10.6	-16.7
	CQL	-2.8	10.5	-9.5	213.3	52.9
CAVI	DQN	8.9	-9.3	-24.6	-48.0	-12.2
	CQL	-8.8	0.5	4.4	1.1	-0.7
AI	DQN	-377.3	-399.3	-386.8	-388.5	-388.0
	CQL	-25.6	235.3	-88.2	-49.9	17.9
SSI	DQN	265.7	-408.2	-334.9	-396.3	-218.4
	CQL	35.8	4.0	11.9	-22.8	7.2

about the nature of the model can help to learn dynamics that better represent reality.

In this work, we built a semi-automated tool that can aid an expert in providing a statistical data-driven validation of her/his intuition about some properties of the environment. Correct deployment of the tool could improve the performance of the optimal policy obtained by solving the learned MDP. Indeed, our results suggest that the proposed algorithm can effectively detect a symmetry of the dynamics of an MDP with high accuracy and that exploiting this knowledge can not only reduce the distributional shift but also provide performance gain in an envisaged optimal control of the system. However, when applied to ORL environments with DNN, all the prescriptions (and issues) about hyperparameter fine-tuning well known to ORL practitioners persist.

Besides its pros, the current work is still constrained by several limitations. We note that the quality

of the approach in continuous MDPs is greatly affected by the architecture of the Normalizing Flow used for Density Estimation and, more generally, by the state-action space preprocessing. In detail, sometimes an environment is endowed by symmetries that an expert can not straightforwardly perceive in the default representation of the state-action space and a transformation would be required (imagine the very same CartPole, but with also the linear speed and position of the car expressed in polar coordinates).

In the future we plan: (i) to expand this approach by trying out more recent Normalizing Flow architectures like FFJORD (Grathwohl et al., 2019); (ii) to consider combinations of multiple symmetries; (iii) after the offline detection of a symmetry, to exploit the data augmentation to improve the learning phase of online agents.

ACKNOWLEDGEMENTS

This work was funded by the Artificial and Natural Intelligence Toulouse Institute (ANITI) - Institut 3iA (ANR-19-PI3A-0004).

REFERENCES

- Angelotti, G., Drougard, N., and Chanel, C. P. C. (2020). Offline learning for planning: A summary. In *Proceedings of the 1st Workshop on Bridging the Gap Between AI Planning and Reinforcement Learning (PRL) at the 30th International Conference on Automated Planning and Scheduling*, pages 153–161.
- Angelotti, G., Drougard, N., and Chanel, C. P. C. (2022). Expert-guided symmetry detection in markov decision processes. In *Proceedings of the 14th International Conference on Agents and Artificial Intelligence - Volume 2: ICAART*, pages 88–98. INSTICC, SciTePress.
- Bellman, R. (1966). Dynamic Programming. *Science*, 153(3731):34–37.
- Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., and Zaremba, W. (2016). OpenAI Gym. *arXiv preprint arXiv:1606.01540*.
- Dean, T. and Givan, R. (1997). Model Minimization in Markov Decision Processes. In *AAAI/IAAI*, pages 106–111.
- Dinh, L., Krueger, D., and Bengio, Y. (2015). NICE: Non-linear Independent Components Estimation. In Bengio, Y. and LeCun, Y., editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Workshop Track Proceedings*.
- Grathwohl, W., Chen, R. T. Q., Bettencourt, J., Sutskever, I., and Duvenaud, D. (2019). FFIORD: Free-Form Continuous Dynamics for Scalable Reversible Generative Models. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.
- Gross, D. J. (1996). The role of symmetry in fundamental physics. *Proceedings of the National Academy of Sciences*, 93(25):14256–14259.
- Kobyzev, I., Prince, S., and Brubaker, M. (2020). Normalizing Flows: An Introduction and Review of Current Methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- 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.
- Levine, S., Kumar, A., Tucker, G., and Fu, J. (2020). Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *ArXiv*, abs/2005.01643.
- Li, L., Walsh, T. J., and Littman, M. L. (2006). Towards a Unified Theory of State Abstraction for MDPs. *ISAIM*, 4:5.
- Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Hiedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. (2015). Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533.
- Narayanamurthy, S. M. and Ravindran, B. (2008). On the Hardness of Finding Symmetries in Markov Decision Processes. In *Proceedings of the 25th international conference on Machine learning*, pages 688–695.
- Paine, T. L., Paduraru, C., Michi, A., Gulcehre, C., Zolna, K., Novikov, A., Wang, Z., and de Freitas, N. (2020). Hyperparameter selection for offline reinforcement learning. *arXiv preprint arXiv:2007.09055*.
- Papamakarios, G., Pavlakou, T., and Murray, I. (2017). Masked autoregressive flow for density estimation. In *Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS’17*, page 2335–2344, Red Hook, NY, USA. Curran Associates Inc.
- Park, D. S., Chan, W., Zhang, Y., Chiu, C.-C., Zoph, B., Cubuk, E. D., and Le, Q. V. (2019). SpecAugment: A simple data augmentation method for automatic speech recognition. In *Proc. Interspeech 2019*, pages 2613–2617.
- Ravindran, B. and Barto, A. G. (2001). Symmetries and Model Minimization in Markov Decision Processes. Technical report, USA.
- Ravindran, B. and Barto, A. G. (2004). Approximate Homomorphisms: A Framework for Non-exact Minimization in Markov Decision Processes.
- Seno, T. and Imai, M. (2021). d3rlpy: An offline deep reinforcement library. In *NeurIPS 2021 Offline Reinforcement Learning Workshop*.
- Shorten, C. and Khoshgoftaar, T. M. (2019). A survey on Image Data Augmentation for Deep Learning. *Journal of Big Data*, 6(1):1–48.
- van der Pol, E., Kipf, T., Oliehoek, F. A., and Welling, M. (2020a). Plannable Approximations to MDP Homomorphisms: Equivariance under Actions. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS ’20*, page 1431–1439, Richland, SC. International Foundation for Autonomous Agents and Multiagent Systems.
- van der Pol, E., Worrall, D., van Hoof, H., Oliehoek, F., and Welling, M. (2020b). MDP Homomorphic Networks: Group Symmetries in Reinforcement Learning. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 4199–4210. Curran Associates, Inc.
- van Dyk, D. A. and Meng, X.-L. (2001). The art of data augmentation. *Journal of Computational and Graphical Statistics*, 10(1):1–50.
- Yarats, D., Brandfonbrener, D., Liu, H., Laskin, M., Abbeel, P., Lazaric, A., and Pinto, L. (2022). Don’t change the algorithm, change the data: Exploratory data for offline reinforcement learning. In *ICLR 2022 Workshop on Generalizable Policy Learning in Physical World*.