Open-Ended Diverse Solution Discovery with Regulated Behavior Patterns for Cross-Domain Adaptation

Kang Xu, Yan Ma, Bingsheng Wei, Wei Li *

Academy for Engineering and Technology, Fudan University, Shanghai, China
kangxu21@m.fudan.edu.cn, {20210860024, weibingsheng, fd_liwei}@fudan.edu.cn

Abstract

While Reinforcement Learning can achieve impressive results for complex tasks, the learned policies are generally prone to fail in downstream tasks with even minor model mismatch or unexpected perturbations. Recent works have demonstrated that a policy population with diverse behavior characteristics can generalize to downstream environments with various discrepancies. However, such policies might result in catastrophic damage during the deployment in practical scenarios like real-world systems due to the unrestricted behaviors of trained policies. Furthermore, training diverse policies without regulation of the behavior can result in inadequate feasible policies for extrapolating to a wide range of test conditions with dynamics shifts. In this work, we aim to train diverse policies under the regularization of the behavior patterns. We motivate our paradigm by observing the inverse dynamics in the environment with partial state information and propose Diversity in Regulation (DiR) training diverse policies with regulated behaviors to discover desired patterns that benefit the generalization. Considerable empirical results on various variations of different environments indicate that our method attains improvements over other diversity-driven counterparts.

Introduction

Deep Reinforcement Learning has exhibited wide success in solving complex tasks, including vision-based video games (Mnih et al. 2015; Jaderberg et al. 2019), quadruped locomotion (Hwangbo et al. 2019; Lee et al. 2020a), and robotic manipulation (Andrychowicz et al. 2020). However, the policies trained in the source environments are prone to fail in environmental variations. For example, dynamics change, such as the damaged component of a robot or encountering a new terrain, might lead to a failure due to poor generalization of the agents. Thus one significant challenge for real-world deployment of RL is the generalization across various conditions.

One natural approach is to train a policy under a range of dynamics in simulation (Tobin et al. 2017; Peng et al. 2018; Rajeswaran et al. 2016), and assume that the trained policy can generalize to the specific target environment. These methods require expert knowledge to manually specify the distribution of training environments in a trial-and-error manner to involve the properties of the target environment. In addition, the trained policy may appear to be over-conservative due to the uncertainty in the training environments (Yu, Liu, and Turk 2018; Xie et al. 2021). Another category of methods trains the policy to implicitly identify the dynamics of the target environment based on samples collected in the target environment during training and then encourages the policy to act optimally according to the identified dynamics (Muratore et al. 2021; Du et al. 2021; Evans, Thankaraj, and Pinto 2022). However, rollouts in the target environment, like the real-world system during training, might result in catastrophic damage due to the premature behavior of the policy.

Recent works have demonstrated that diversity-driven policies can extrapolate to new environments through the few-shot adaptation (Eysenbach et al. 2018; Kumar et al. 2020a; Osa, Tangkaratt, and Sugiyama 2021; Parker-Holder et al. 2020; Zhou et al. 2022). While the policy population with different behavior characteristics can generalize to different environment variations, the learned policies may result in potential safety problems in practical scenarios like real-world systems, as the behaviors of the diverse policies are unpredictable. Especially, the degree of diversity that is necessary for the generalization may be limited. For instance, when we aim to obtain multiple policies for a quadruped robot that can generalize to various terrains, what we desire might be policies with different locomotion patterns rather than those able to roll on the ground. However, these works train the policies without regularizing the behavior, which might result in inadequate feasible policies.

In this work, we take the first step towards diverse policies with regulated behaviors for generalization. To encourage sufficient feasible solutions for adaptation in a wide range of downstream scenarios, we propose a novel diversity objective based on the divergence of inverse dynamics models $T^*(a|f(s), f(s'))$ under partial state information. The partial state information is removed by utilizing a customizable state filtration function $f(s)$. Intuitively, the actions impacting the remaining state information would be discouraged from getting diversified, thus regulating the behaviors of trained policies. Additionally, we introduce the...
open-ended training manner to achieve continuous solution discovery, which avoids the drawback of prior work training with a fixed number of policies (Kumar et al. 2020; Parker-Holder et al. 2020).

The main contribution of our work is the proposal of a diversity-driven algorithm, Diversity in Regulation (DiR), which trains multiple policies with regulated behavior patterns for efficient generalization by diversifying the action distributions in a customizable way. Our analysis demonstrates that the discovered policies show more regulated behaviors against prior diversity-driven approaches, which benefits generalization across a wide range of test conditions. Empirically, we observe that DiR substantially outperforms prior methods under various environment discrepancies.

Preliminaries

Notation. To model the sequential decision problem, we consider the standard Markov Decision Process (MDP) (Sutton and Barto 2018) as \((S, \mathcal{A}, \mathcal{P}, r, \mu, \gamma)\), where \(S\) and \(\mathcal{A}\) are the state space and action space respectively; \(\mathcal{P}(s'|s, a) : S \times \mathcal{A} \times S \rightarrow [0, 1]\) defines the dynamics of the environment and specifies the transition probability of reaching \(s'\) at the next step given current state \(s\) and the executed action \(a\); \(r(s, a) : S \times \mathcal{A} \rightarrow [0, 1]\) denotes the reward function; \(\mu(s) : S \rightarrow [0, 1]\) denotes the distribution of initial states and \(\gamma \in (0, 1)\) is the discount factor. Considering a policy \(\pi(a|s) : S \times \mathcal{A} \rightarrow [0, 1]\) which outputs the probability of choosing action \(a\) given the state \(s\), the probability density function of any trajectory \(\tau = \{s_0, a_0, s_1, a_1, \ldots, s_T\}\) can be formulated as

\[
P(\tau) = \mu(s_0) \prod_{t=0}^{T-1} \pi(a_t|s_t) \mathcal{P}(s_{t+1}|s_t, a_t).
\]

Inverse dynamics. Here we denote the inverse dynamics of the MDP as \(T(a|s, s') : S \times S \times \mathcal{A} \rightarrow [0, 1]\) which defines the probability of action \(a\) given the state pair \((s, s')\) in the consecutive steps. Since there might be various actions under different policies given the state pair \((s, s')\), the inverse dynamics under some specific policy \(\pi\) can be formulated as:

\[
T^\pi(a|s, s') = \frac{\mathcal{P}(s'|s, a)\pi(a|s)}{\mathcal{P}(s'|s, \bar{a})\pi(\bar{a}|s)d\bar{a}}.
\] (1)

State filtration function. In this work, we assume a customizable function \(f(s) : S \rightarrow \hat{S}\) that removes partial state information from the state \(s\). For instance, \(f(s)\) can be defined to remove the x-axis coordinate in a navigation task whose full state information includes 2D coordinates. The resulting partial state \(\hat{S}\) can also be considered as the state space from a partially observable MDP (POMDP) (Bellman 1957).

Mutual-Information in RL. Mutual information can be generally expressed as

\[
I(X; Y) = \int \int p(x, y) \log \frac{p(x, y)}{p(x)p(y)} dy dx = H(X) - H(X|Y) = H(Y) - H(Y|X),
\] (2)

which defines the mutual dependence between two random variables. Mutual information has been introduced to find the best representation subject to a constraint on the complexity (Tishby, Pereira, and Bialek 2000; Alemi et al. 2016). In the context of RL, maximizing the mutual information \(I(S; Z) = H(Z) - H(Z|S)\) between visited states \(S\) and latent variable \(Z\) has been proposed to discover diverse policies (Eysenbach et al. 2018; Kumar et al. 2020).

Diverse high-performing policies. Prior approaches that train multiple high-performing policies \(\{\pi_{\theta_k}\}_{k=1}^M\) with diverse behaviors\(^1\) can be naturally formulated as:

\[
\max_{\pi_k} \mathbb{E}_{\pi_k} \left[ \sum_{t=1}^{T} \gamma^{t-1} r_t \right], \quad 1 \leq k \leq M \tag{3}
\]

s.t. \(D(\pi_i, \pi_j) \geq \delta, \quad \forall 1 \leq i, j \leq M, i \neq j\),

where \(D(\cdot, \cdot)\) measures how different the two policies are, and \(\delta\) is the diversity threshold. One natural choice of the distance measurement function is the KL divergence \(D_{KL}(\pi_i|\cdot || \pi_j|\cdot)\) which is widely adopted in prior works (Schulman et al. 2015, 2017; Hong et al. 2018).

Few-shot adaptation. In novel environments, we rollout each policy from the trained population for several episodes and deploy the best-performing one, which resembles prior diversity-driven approaches (Kumar et al. 2020; Osa, Tangkaratt, and Sugiyama 2021; Derek and Isola 2021).

Motivation Example: Inference Variance of Inverse Dynamics Models

We motivate our method with an empirical observation about the inference variance of the inverse dynamics under different state information settings. Here we train four independent inverse dynamics models (IDMs) \(\{T_{\theta_i}(a|s, s')\}_{i=1}^4\) simultaneously using different state settings in Walker from Mujoco (Todorov, Erez, and Tassa 2012). The training data are collected by an agent trained

\(^1\)The \(\theta\) will be omitted for simplicity.
through Soft Actor-Critic (SAC) (Haarnoja et al. 2018). We
utilized four different state settings for the inverse dynamics
models with varying degrees of information missing. We
train the inverse dynamics models by maximizing the log-
likelihood:

\[ L(\phi_i) = E_{(s,a,s')} \sim D} [\log T_{\phi_i}(a|f_i(s), f_i(s'))], 1 \leq i \leq 4, \]

where the \( f_i(s) : S \rightarrow \tilde{S} \) is the state filtration function
corresponding to the model \( i \) that removes some specific
state information. After training, we evaluate the inference
variance \( \text{Var}[T_{\phi_i}(\cdot|s, s')] \) of the IDMs given the same batch
of data \( D_{\text{test}} = \{(s, s')\} \), and the results are shown in
Figure 1. The results indicate that the variance of the inverse
dynamics model increases with the increase of missing state information. The details of the experiment refer to Appendix A.

Herein we interpret the observation with a toy example
of a 2-DOF robotic arm, as shown in Figure 1. When
we employ the joint angle \( \theta_1 \) as a partial state space
\( \tilde{S} \), there might be multiple possible actions given any
state pair \((\tilde{s}, \tilde{s}')\). In contrast, there might be only one
possible action given the state pair \((s, s')\) under the full
state information setting. Motivated by the observation, we
propose to maximize the divergence of IDMs with the partial
state setting under different policies to encourage them to
produce distinct action distributions given any \((f(s), f(s'))\)
pair. Furthermore, we introduce customizable state filtration
functions to specify the desired patterns. Semantically, the
removing state information (e.g., \( \theta_2 \) above) can be regarded
as the state information of body parts that is unnecessary
to be diversified (e.g., roll angles of a quadruped robot).
In contrast, we can remove the state information like leg
motions to obtain policies with diverse locomotion patterns.
Thus, we introduce the state filtration function to focus on
discovering desired behaviors.

We denote the divergence between inverse dynamics under
the partial state information setting as
\( P^{\pi_j}(a|f(s), f(s')) : \tilde{S} \times \tilde{S} \times A \rightarrow [0, 1] \), and the overall
objective can be formulated as the following constrained
optimization problem:

\[
\begin{align*}
\max_{\pi_k} E_{\pi_k} \left[ \sum_{t=1}^{T} \gamma^{t-1} r_t \right], & \quad \forall 1 \leq k \leq M \quad (5) \\
\text{s.t. } E \left[ D_{KL} [P^{\pi_j}(|f(s), f(s')) \| P^{\pi_k}(|f(s), f(s'))] \right] \geq \delta, & \quad \forall 1 \leq i, j \leq M, i \neq j,
\end{align*}
\]

where \( f(s) : S \rightarrow \tilde{S} \) removes specific partial state information from \( s \), and can be designed to control the diverse patterns we aim to discover.

\section*{Diversity in Regulation}

This section presents our approach to resolving the objective
in detail. We propose an open-ended training manner for
diverse solution discovery, transform the objective into a
trivial form, and finally analyze the connection between
our method and prior diversity-driven approaches through
mutual information. The overview of DiR is shown in Figure 2.

\begin{figure}
\centering
\includegraphics[width=\textwidth]{fig2.png}
\caption{The semantic overview of DiR. \( \pi_i \) and \( \pi_j \) represent the policy and inverse dynamics model, respectively. We train multiple policies in the source environments in an iterative manner and generalize to the variation condition with the best-performing one after the few-shot adaptation.}
\end{figure}

\section*{Open-ended Solution Discovery}

Directly solving Eq. 3 or Eq. 5 requires the non-trivial
pairwise constraints computation for training each policy
and the parallel framework for the policy population, which
limits the population size in prior works (Parker-Holder et al.
2020; Masood and Doshi-Velez 2019). Here we introduce an
iterative training manner that trains only a single policy at
a time and optimizes the policy \( \pi_k \) to be distinct from the
previously discovered policies \{\( \pi_i \}_{i=1}^{k-1} \}, which resembles
the prior work (Zhou et al. 2022). However, we concentrate
on the regulated diversity objective. Formally, we optimize
the following objective to train a policy \( \pi_k \) at each iteration:

\[
\begin{align*}
\max_{\pi_k} E_{\pi_k} \left[ \sum_{t=1}^{T} \gamma^{t-1} r_t \right] \quad (6) \\
\text{s.t. } E \left[ D_{KL} [P^{\pi_j}(|f(s), f(s')) \| P^{\pi_k}(|f(s), f(s'))] \right] \geq \delta, & \quad \forall 1 \leq i < k,
\end{align*}
\]

which converts the optimization of the whole population
simultaneously to the optimization of a single policy with a
simplified constraint. Furthermore, we can discover policies
exhaustively through open-ended training to obtain some
specific behavior pattern we desire, which is superior to prior
works that fix the number of policies.

\section*{Diversity via Inverse Dynamics Disagreement}

To solve the constrained optimization problem in Eq. 6, we
introduce the Lagrangian multiplier method to convert the
hard constraints to soft penalties:

\[
\begin{align*}
\max_{\pi_k} E_{\pi_k} \left[ \sum_{t=1}^{T} \gamma^{t-1} r_t \right] + \sum_{i=1}^{k-1} \beta_i \text{Div}(\pi_k, \pi_i), \quad (7) \\
\text{where } \text{Div}(\pi_k, \pi_i) = E_{\pi_k} \left[ D_{KL} [P^{\pi_k}(|f(s), f(s')) \| P^{\pi_i}(|f(s), f(s'))] \right],
\end{align*}
\]

where \{\( \beta_i \}_{i=1}^{k-1} \} are the multipliers that can be considered
as hyperparameters, and \( \text{Div}(\pi_k, \pi_i) \) can be interpreted
as the inverse dynamics disagreement between two policies. Introducing the Lagrangian multipliers method to simplify the constrained optimization problem and set the multipliers as hyperparameters is widely used in the RL research (Stooke, Achiam, and Abbeel 2020; Chanesane, Schmid, and Laptev 2021; Peng et al. 2018). While the constraints might exhibit oscillations during training (Stooke, Achiam, and Abbeel 2020), it is acceptable since we aim to encourage diversity rather than obtain severely distinct policies.

To tractably optimize the $Div_{i,k}(π_k, π_i)$, we introduce an ensemble of inverse dynamics models $\{T_{\phi_k}\}_{k=1}^M$ to approximate the inverse dynamics under corresponding policies. At each iteration, we train the inverse dynamics model $T_{\phi_k}$ simultaneously by maximizing the log-likelihood:

$$L(\phi_k) = E_{π_k} \left[ \log T_{\phi_k}(a|f(s), f(s')) \right].$$

(8)

To solve the optimization for policy $π_k$ in Eq. 7, we approximate the $P_{π_k}(a|f(s), f(s'))$ with an inverse dynamics model $T_{\phi_k}(a|f(s), f(s'))$, and thus convert the diversity objective as follows:

$$DivEC_{i,k}(π_k, π_i) = E_{π_k} \left[ -\log T_{\phi_k}(a|f(s), f(s')) \right],$$

(9)

which can be interpreted as the cross-entropy between the transitions collected by policy $π_k$ and the inverse dynamics of $π_i$. We present in Appendix B.1 that the diversity objective in Eq. 9 can be approximately lower bounded by maximizing the divergence between the inference action $f(s)$ and $f(s')$ with the proposed divergence determinant, vanilla PPO with the population size $10^5$, and IDMs. We compare DiR to several diversity-driven approaches through the state filtration function. (3) Does our method discover diverse policies with regulated behavior patterns through the state filtration function? (3) Since we hypothesize DiR can obtain more diverse policies compared with baselines given the same population size, can the trained population perform better in a wide range of dynamics mismatch scenarios? Implementation details and additional results are presented in Appendix.

**Experiment**

In this section, we aim to empirically answer the following questions: (1) Can our method discover diverse behaviors? (2) Does our method discover diverse policies with regulated behavior patterns through the state filtration function? (3) Several diversity-driven approaches can regulate the diverse behaviors to a specific coverage of patterns by customizing the state filtration function.

### Experimental Settings

**Environments.** We adopt four continuous control tasks: Ant, Walker, Hopper, and Minitaur from Mujoco (Todorov, Erez, and Tassa 2012) and Bullet (Coumans and Bai 2016–2019), as illustrated in Appendix. C. We implement extensive test scenarios, including broken leg joints, shifted dynamics parameters, and sensor failure conditions. See Appendix C.1 for details of the environments.

**Customized state filtration functions.** We focus on discovering diverse locomotion patterns (e.g., walking with different legs in Ant) by designing the state filtration function for the locomotion tasks. Thus we remove partial state information about the leg motions (e.g., joint positions) through $f(s)$ in all four environments. Full details of the state filtration functions and the original state information are described in Appendix C.2.

**Baselines and implementation.** We compare DiR to SMERL (Kumar et al. 2020) that trains diverse policies for generalization to environmental variations, DvD (Parker-Holder et al. 2020) that trains an ensemble of policies via the proposed divergence determinant, vanilla PPO with multiple independent policies (Multi), vanilla PPO with a single policy (PG). We set the population size as 10 in all
baselines, and all baselines except PG train the same number of policies. We train each policy with 2M steps for all algorithms. Each trial runs eight times with different random seeds. Refer to Appendix C.3 for implementation details.

Emergent Behavior with DiR
Since we remove the state information of one leg in Walker, we hypothesize that our method can learn various locomotion patterns (e.g., hopping, walking) with different legs. Thus, we visualize the foot contact within an episode in Walker, as shown in Figure 3. The results suggest that our method can iteratively discover diverse policies with different locomotion patterns, including hopping on both feet, incomplete one-leg hopping, and complete one-leg hopping. Furthermore, there is no significant performance degradation resulting from diversity-driven training.

Population Comparison in Training Environments
Herein we aim to compare the policies with prior diversity-driven approaches quantitatively. We adopt the population diversity, a determinant-based diversity paradigm proposed in (Parker-Holder et al. 2020), to quantify the behavior diversity of the trained policies. Here we collect 2000 states for the behavior embeddings in each environment. The results are reported in Table 1, where DiR outperforms all baseline methods concerning the diversity score. We observe that the independently trained policies in Multi can also obtain competitive diversity scores compared to DvD and SMERL which introduce extra diversity objectives, which may result from the different initialization of the policies as presented in (Jiang and Lu 2021).

To examine whether DiR can train more practical locomotion patterns by inducing the regulated diversity objective, we adopt the approach that describes behaviors of quadruped robots in prior works (Cully et al. 2015; Nilsson and Cully 2021), which computes the time proportion when the feet contact the ground within an episode. The visualization of the behavior diversity in Walker and Ant is shown in Figure 4. The results show that the behaviors of each policy trained by DiR are more consistent in multiple episodes, and DiR can discover more distinct locomotion patterns compared with baselines.

Figure 4: Visualization of behaviors through the time proportion when feet contact the ground, where the different colors indicate different policies. Different colors represent different policies from the population, and we run each policy for 20 episodes. (Top): Time proportion when two feet contact the ground within an episode in Walker. (Bottom): Time proportion when two adjacent feet contact the ground within an episode in Ant.

Furthermore, here we show the percentile performance of the population, as shown in Figure 5. The results indicate that DiR achieves competitive performance in Hopper and Minitaur while obtaining better averaged and worst-case performance in Walker and Ant, compared with baselines. We believe that regulated diversity can hinder the discovery of behavior patterns with poor performance, which results in improved worst-case performance over the policies.

Adaptation in Environment Variations
To examine whether DiR can provide adequate feasible policies which benefit the generalization, we implement various variations of the environments, including the crippled legs, the shifts of the dynamics parameters (e.g., mass), and sensor failures. See Appendix C.1 for detailed descriptions of the test conditions. For few-shot adaptation,
Figure 5: Percentile performance of the policy population across all four environments. The 0 percentile on the x-axis represents the worst-performing policy, while the 100 percentile represents the best-performing one.

Table 1: Diversity scores across all environments. Asterisks indicate that the results are significantly different from all the baselines under p < 0.05.

<table>
<thead>
<tr>
<th></th>
<th>Ant</th>
<th>Hopper</th>
<th>Walker</th>
<th>Minitaur</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multi</td>
<td>71.6(0.4)</td>
<td>57.8(2.4)</td>
<td>76.4(0.4)</td>
<td>81.6(0.2)</td>
</tr>
<tr>
<td>DvD</td>
<td>71.6(0.5)</td>
<td>59.1(2.2)</td>
<td>76.2(0.4)</td>
<td>81.4(0.2)</td>
</tr>
<tr>
<td>SMERL</td>
<td>71.5(0.5)</td>
<td>63.2(2.4)</td>
<td>77.1(0.4)</td>
<td>81.7(0.3)</td>
</tr>
<tr>
<td>DiR</td>
<td><strong>73.0(0.4)</strong></td>
<td><strong>66.1(1.7)</strong></td>
<td><strong>78.4(0.8)</strong></td>
<td><strong>83.2(0.3)</strong></td>
</tr>
</tbody>
</table>

Figure 6: Adaptation performance under different levels of the dynamics parameter variations.

we run each policy in the environment for 20 episodes and report the performance of the best-performing one.

We first evaluate the approaches under conditions with damaged body components. As the results show in Table 2, DiR outperforms baseline methods on most test conditions. In the variations of Hopper, DiR achieves comparable performance with baselines, which can result from the morphology with only one leg that limits the possible locomotion patterns. However, given the other three environments where the robots are multi-legged, DiR surpasses baselines thanks to the regulated diversity discovery. As we utilize the state filtration functions that remove the state information about the legs, DiR will focus on discovering policies that behave differently in terms of the leg motions. Thus, DiR has sufficient strategies (e.g., walking on a single leg) to handle the situations like the damaged leg. Specifically, DiR is the only approach that adapts with better locomotion patterns in the test Walker environment where a foot joint is broken. Furthermore, we remark that PG training a single policy performs significantly worse than the methods training multiple policies, which indicates that the diversity-driven approaches are simple yet effective for adaptation under dynamics variations. Importantly, we verify that the superior adaptation performance of DiR results from different policies, which further validates that the diverse behaviors benefit the extrapolation to different environments. The details of selected policies in the test environments are presented in Appendix D.1.

Furthermore, we consider the adaptation to environments with shifted dynamics parameters. Here we scale the ankle friction or the leg mass of Ant, and the foot friction or the foot mass of Hopper and Walker. As the adaptation performance in Figure 6 shows, DiR outperforms baselines in Walker and Ant, while DiR produces comparable results with baselines in Hopper. Specifically, DiR provides a non-trivial improvement in Walker - foot friction and Ant - leg mass compared to all baselines, where the performance does not drop significantly when the value of foot friction gradually increases. Furthermore, we observe that DiR achieves better performance in the environment where the dynamics parameter is the same as the training environment (scale = 1), which can result from the stationary diversity-driven intrinsic rewards provided by the fixed inverse dynamics models converged in early iterations. In contrast, DvD and SMERL, whose intrinsic rewards are non-stationary due to the simultaneously trained population or discriminators, might cause unstable training dynamics and thus result in performance degradation. The phenomenon further validates the advantage of the open-ended training manner.

Finally, we implement test environments with various sensor failures where the corresponding state variables are always zeros and report the adaptation performance of the policies. The results in sensor failure conditions of Ant are shown in Table 3, where we observe that DiR is weaker than baselines when the sensors on leg 2 are defective, which might result from reliance on the state information of Leg 2 for the decision making. However, DiR outperforms baseline methods in most environments, which validates that the policies trained through DiR output actions conditioning
on different state variables of the state information and further verify the robustness of DiR. Additional results in Walker are shown in Appendix D.2.

### Related Work

In this work, we focus on the generalization across environments with various dynamics. A common approach to solve this problem is through domain randomization (Tobin et al. 2017; Peng et al. 2018), where a single policy is trained under various dynamics in simulation. Prior works have shown the effectiveness of domain randomization for the adaptation across dynamics (Rajeswaran et al. 2016; Yu et al. 2017; Akkaya et al. 2019; Shi et al. 2022; Mehta et al. 2020). Another line of work resolves the generalization through domain adaptation (Chebotar et al. 2019; Hwangbo et al. 2019; Ramos, Possas, and Fox 2019), which grounds the simulator with the collected transitions from the target domain and train the policy to be optimal under the target dynamics. Furthermore, Robust RL has shown improved transfer performance by optimizing the worst-case performance in the source environment (Pinto et al. 2017; Jiang et al. 2021; Mankowitz et al. 2019). In contrast, we resolve the generalization using an ensemble of diverse policies.

Searching for diverse solutions has been studied in Evolutionary Computation and Reinforcement Learning research. In Evolutionary Computation, Quality-Diversity (QD) is a representative type of approach that searches for diverse high-performing solutions (Cully et al. 2015; Mouret and Clune 2015; Nilsson and Cully 2021). However, the requirement of defining behavior descriptors limits QD to complicated tasks (Grillotti and Cully 2022). In Reinforcement Learning, unsupervised skill discovery has been proposed to train a latent conditional policy without environment reward (Eysenbach et al. 2018; Sharma et al. 2020; Hartikainen et al. 2019), which can prevent the behavior from being practically feasible by ignoring the environment reward. When extrinsic rewards are also considered, several approaches have been proposed to train diverse high-performing policies (Kumar et al. 2020; Parker-Holder et al. 2020; Masood and Doshi-Velez 2019; Zhou et al. 2022; Lupu et al. 2021; Zahavy et al. 2021; Zhang, Yu, and Turk 2019). We remark that our open-ended training manner resembles the prior works that train diverse policies iteratively (Zhou et al. 2022; Zhang, Yu, and Turk 2019). However, we focus on controllable diversity through the exhaustive solution discovery different from the methods. Furthermore, several approaches resolve the generalization over various dynamics with the assistance of diverse policies (Kumar et al. 2020; Osa, Tangkaratt, and Sugiyama 2021; Kaushik, Arndt, and Kyrki 2022), same as our work. However, we take the first step to regulate the diversity for more efficient adaptation as far as we know.

Our proposed diversity optimization through the inverse dynamics disagreement also resembles Model-based RL (MBRL) (Deisenroth and Rasmussen 2019; Chua et al. 2018). The divergence between inverse dynamics models has also been proposed in the prior work for imitation learning from the observation (Yang et al. 2019). For generalization across dynamics, recent work has achieved the generalization of the dynamics model (Lee et al. 2020b; Seo et al. 2020). Unlike these works, we utilize the inverse dynamics models for diverse solution discovery.

### Conclusion

In this work, we present Diversity in Regulation (DiR), a novel diversity-driven algorithm that learns multiple high-performing policies iteratively for adaptation under dynamics variations. The key ingredient of our method is the novel diversity objective through the inverse dynamics disagreement with the state filtration function. Specifically, we can regulate the diversity by customizing the state filtration function for desired behavior patterns. Our empirical results show that DiR can adapt to various test conditions and outperforms prior diversity-driven approaches. Overall, we believe our approach would further strengthen the understanding of diverse solution discovery and could be helpful in safe adaptation under dynamics variations which is critical for the Sim2Real problem.
Acknowledgments
This research was supported in part by Shanghai Municipal Science and Technology Major Project (No.2021SHZDZX0103), in part by Ji Hua Laboratory, Foshan, China (No.X190011TB190), in part by Science and Technology Development Center, Ministry of Education (No.2021ITFA10013).

References


