Isolating Stochastic Sources of Policy Divergence in Proximal Policy Optimization

Anonymous authors
Paper under double-blind review

Abstract

This work presents an analysis of how Reinforcement Learning (RL) policies diverge during training, aiming to separate and investigate divergence arising from environmental randomness and from the policy's initialization. We trained 300 policies using Proximal Policy Optimization (PPO) across three contexts in the MinAtar testbed: (1) with fixed parameter initialization, (2) with fixed sampled scenarios, and (3) with neither fixed. The resulting policies were examined for performance, feature-attribution overlaps, action disaccord, and overlaps in critical state estimates. The distributions of policies are similar across all training contexts for all examinations, except that the overlap in feature attributions increases when the initial parameters are fixed. These results show that, despite controlling for parameter initialization and the scenarios drawn from the environment, the PPO policies diverge similarly across training contexts. The results, therefore, suggest that *PPO exhibits severe path dependence*: the unpredictability of the final policy is deeply ingrained in PPO's stochastic exploration-update loop. Furthermore, our investigation demonstrates that similarly trained PPO policies exhibited substantial differences in *how* they solved the RL tasks.

1 Introduction

Sequential decision-making tasks, such as controlling a robot or a vehicle, involve one or more agents making a series of choices over time, with each choice influencing the evolving situation. These tasks are often executed in environments with vast operational spaces, requiring the agent to gather and process high-dimensional sensory data, such as images, and to plan across long horizons. Reinforcement Learning (RL) (Sutton & Barto, 2015; Morales, 2020) algorithms only need a reward signal to learn from experience, both how to process inputs and how to act, and have demonstrated high performance in games (Silver et al., 2016; Young & Tian, 2019), control tasks (Todorov et al., 2012), and industrial applications (Ramos & Busboom, 2025; Panzer & Bender, 2022).

Despite these successes, RL training remains inherently unstable. These algorithms must balance feature learning, environmental exploration, and exploitation of gathered experience, with complex interdependencies that render them highly path-dependent; similar optimization runs may converge to distinct optima. Consequently, they exhibit high sensitivity to hyperparameters (Henderson et al., 2018) and random seeds (Bjorck et al., 2021; Jang et al., 2022; 2025), meaning that marginal modifications in the implementation may produce major and erratic differences in the final policy. Such unpredictability creates a significant consistency gap (Islam et al., 2017; Agarwal et al., 2021; Henderson et al., 2018), undermining practitioner trust (Chan et al., 2019) and challenging the scientific validity of reported benchmarks.

Several works have proposed evaluation strategies (Agarwal et al., 2021; Henderson et al., 2018; Chan et al., 2019) to narrow this consistency gap, with Chan et al. (2019) highlighting that evaluation should also account for the risk of a failed training run, i.e., a training run in which the agent fails to learn early and never recovers. Furthermore, Engstrom et al. (2020) argued that many of the performance gains attributed to primary algorithmic innovations, such as the Proximal Policy Optimization (PPO) objective function (Schulman et al., 2017), could be a consequence of secondary 'code-level' optimizations, such as annealing the learning rate. Others have proposed methods to reduce the performance variability of RL, such as entropy-aware

model initialization (Jang et al., 2022) to facilitate initial exploration and normalizing penultimate-layer features (Bjorck et al., 2021) to mitigate a vanishing-gradient problem arising from information loss of the state in the feature representation. Similarly, Moalla et al. (2024) examine the degradation of learning outcomes from new experiences in PPO agents as the training process advances, termed *capacity loss*, identify a similar vanishing-gradient problem as the cause, and thus propose regularization of the penultimate-layer feature representations as a remedy. Recently, Jang et al. (2025) conducted an empirical study of PPO agents, investigating the hypothesis that divergent exploration among individual agents during training is a key contributor to the performance variability. Their analysis suggests that in environments where individual agents diverged more in their exploration, they would indeed vary more in their performance.

In this work, we aim to contribute to understanding *how* RL policies diverge across training runs. To this end, we conduct a controlled empirical experiment in which we fix part of the randomness in the training algorithm to investigate how this affects the distribution of trained policies. Specifically, to investigate how much the randomness in scenario sampling from the environment shapes policies during training, we fix the network model's initial weights; we refer to this training context as *environmental divergence*. Vice versa, to isolate variation in final policies arising from varying the initial weights of the network models, we fix the seeding of the random number generator in the environment, such that it samples the same initial states and is seeded similarly in the concurrent transitions across all policies trained in this context (i.e., all non-agent entities have fixed trajectories), but the individual policies are free to diverge based on their action choices, and refer to this context as *algorithmic divergence*. In the third training context, we fix neither, as a *baseline*. See Figure 1 for a high-level illustration of these training contexts. We train 300 PPO policies in each training context for each environment in the MinAtar testbed (Young & Tian, 2019), enabling a comparison of the final policy distributions across varying levels of environmental complexity and difficulty.

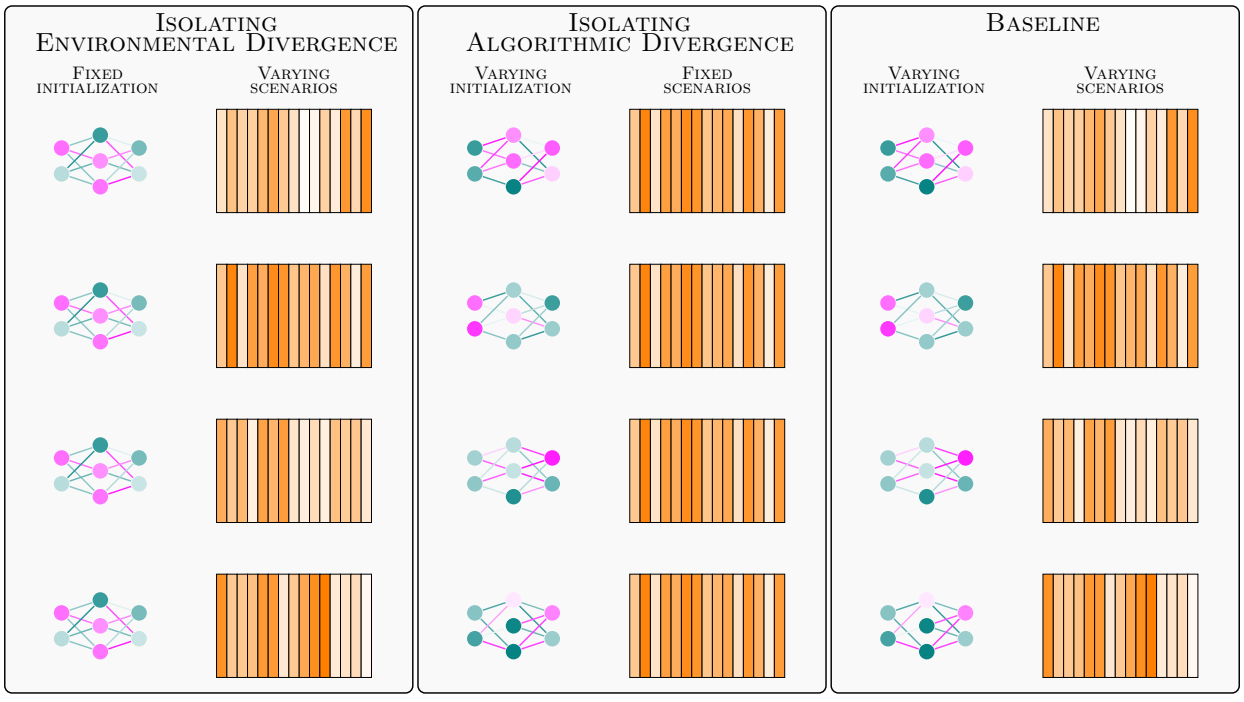

Figure 1: A high-level visualization of the three training contexts, where the colors on the miniature neural networks represent the values of the initial parameter values, and each rectangle's gold shade color represents a different scenario, or level, sampled from the environment for training. In the first training context, we seek to isolate policy variation arising from environmental sampling randomness by training the same network with the same initial weights on different sets of scenarios sampled from the environment. In the second, we seek to isolate the divergence in resulting policies solely due to randomness in parameter initialization by training with varying parameter initializations on the same set of environment scenarios. In the third training context, as a baseline, we vary both.

Whereas previous work emphasizes failed training runs, our analysis focuses on mapping the *divergence* between similarly trained policies, i.e., on *how* they differ. In addition to investigating performance variation across the three training contexts, we examine policies along four axes. First, we investigate the *complementarity* of policy sets by evaluating a majority-vote ensemble of these sets. From Wood et al. (2023), we know that ensemble improvement is related to *prediction diversity*, i.e., disagreement on the prediction values among the individuals of the ensemble. However, because multiple valid RL solutions may exist, the lack of a common ground truth can complicate ensemble improvement; therefore, we refer to this as an investigation of complementarity rather than prediction diversity. Second, to examine variation in *how policies interpret the state* when choosing actions, we present the alignment of policies of their most salient features in the observation. Third, as a proxy for investigations into strategic diversity, we present *disaccord in action choices* within the policy sets. Finally, to investigate how the policies align in their *assessments of state criticality*, i.e., the degree to which choosing the appropriate action in a given state is crucial, we present the distributions of intersections among the most important states within the policy sets.

Our empirical analysis indicates that the three training contexts yield similar performance distributions, similar diversity in action choices, and similar overlap distributions for the states they identify as important to the episodes' success. We also found that the policies aligned substantially more in their feature attributions when the initial policy parameters were fixed compared to the other two training contexts. This reveals a decoupling between state representation and strategic execution: even when weight initialization biases agents to interpret the environment similarly, their operational strategies diverge as severely as in fully uncontrolled baselines. We demonstrate that fixing these initial conditions of the PPO algorithm is insufficient to constrain policy behavior, as the highly sensitive feedback loop between stochastic exploration and policy updates drives the solutions toward a multiplicity of distinct yet viable strategic optima. Our findings underscore the critical need for large-scale evaluation protocols to assess the reproducibility of DRL.

2 Background

The Reinforcement Learning Problem: RL problems are modeled as *Markov Decision Processes* (MDPs) Sutton & Barto (2015), with defined state and action spaces (denoted $\mathcal{S}$, and $\mathcal{A}$, respectively), where the transition to a new state is determined solely by the choice of action and the previous state, through a (possibly stochastic) transition function, $T : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$, and each transition is coupled with a reward signal, $R : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$. The goal is to learn, through trial and error, a policy, $\pi : \mathcal{S} \to \mathcal{A}$, that maximizes the expected cumulative rewards. In *Deep Reinforcement Learning*, the policy is derived from a model that can learn functions over a high-dimensional input space, such as a *neural network*.

Ensemble Methods: In the supervised learning regime, Wood et al. (2023) developed the bias-variance-diversity tradeoff for a wide range of loss functions. They formulate ensemble learning as a trade-off among the average individual biases, the average individual variances, and the disagreement among individual predictions within the ensemble, which they call diversity. Specifically, ensemble diversity reduces the expected risk beyond that of the individual models' average risk. However, the additional strategic diversity among viable RL solutions further complicates the theory behind ensemble improvement in the RL paradigm, as there may be no common ground truth. A common hypothesis for the success of ensembling in RL is that policies fail in different states (Wiering & Van Hasselt, 2008), and thus, when combined, the majority vote ensemble typically masks individual faults.

Critical States: The notion of a *critical state* is defined as a state where the choice of action is *crucial to the episode's success* (Huang et al., 2018; Amir & Amir, 2018; Spielberg & Azaria, 2019). Because policies must be reliable in critical states, they are of high interest to developers and researchers. However, identifying critical states in complex sequential decision-making tasks is not straightforward, and we often use the policy's *state importance* measure to guide our search. For PPO policies, we link the state importance measure to the uncertainty of the action distribution at a state, as in Huang et al. (2018), using *policy confidence as a proxy for deemed state importance*. The uncertainty of the policy can be quantified as the

entropy of the action distribution as

$$H[\pi(\cdot|s)] = -\sum_{a \in \mathcal{A}} \pi(a|s) \ln \pi(a|s),$$

where $\pi(\cdot|s)$ is the action distribution of the policy π at state s. The state's importance increases as the entropy of the action distribution decreases; we use this measure to rank states by criticality based on the policy's perception.

3 Study Design

Our study is designed to investigate the role of stochastic sources in the divergence of trained policies across multiple aspects: the performance of individual policies, the performance of ensemble policies of the policy sets, the policy sets' state interpretations, the policy sets' disaccord in action choices, and the policy sets' agreement in state criticality.

We study the divergence of policies in three training setups (see Figure 1): one in which the policy parameter initialization is fixed, one in which the scenarios sampled from the environment are fixed, and one in which both the parameter initialization and the scenarios are randomly sampled. In the first case, we aim to isolate the *algorithmic divergence*, defined as the divergence arising from random parameter initialization. The second training setup, in which the scenarios are fixed, aims to isolate *environmental divergence*, the divergence in trained policies arising from the randomness of scenario sampling. Finally, in the third setup, to have a baseline divergence to compare against, and indeed referred to as the *baseline*, we perform standard training where both parameter initialization and the randomness in sampling from the environment are varied. We repeated all experiments three times with different random seeds to mitigate artifacts arising from any single seed.

Examining the variation in performance: Performance variation is typically highlighted when investigating the instability of RL schemes. Therefore, we investigate performance variability across training contexts. To this end, we evaluate all policies and report the distributions of mean accumulated rewards over 1024 episodes. Furthermore, to investigate how *complementary* the trained policies are across the three training setups, we ensemble 20 best-performing policies from each setup and evaluate the majority-vote ensemble policy. By complementary, we mean *diversity in predictions* in the sense used in Wood et al. (2023) for supervised learning, but it also encompasses all possible sources of improvement in sequential decision-making tasks, such as reducing the probability that the majority makes the same mistake in the same time step.

Investigating the policies' state interpretations: The Rashomon effect in Explainable Machine Learning (Müller et al., 2023) demonstrates that, even with the same training data, multiple high-performing solutions can emerge, possibly driven by the model's considerable freedom in its latent representations. Therefore, to study divergence in interpretations of the state in RL, we investigate how policies interpret states. Specifically, we use the 20 high-performing policies in each training context as a policy set. For each policy set, we extract the top five salient features from the same 2,500,000 states found by executing the majority-vote ensemble of the baseline sets (B), using normal backpropagation on the summed action distributions across all policies in the set. Although the state observations have a relatively large number of possible features (the smallest are $10 \times 10 \times 4 = 400$ in Asterix and Breakout, while the largest is $10 \times 10 \times 10 = 1000$ in Seaquest), each state observation channel is a binary grid with typically only a few features active, effectively giving a smaller number of features to find salient. The divergence in state interpretations is then measured by pairwise overlap in the five most salient features across the policy sets.

Inspecting disaccord in action choices: Strategic solution diversity, referring to the diversity of policies' action choices, whether valid or invalid, is a known feature of trained RL policies (Chapter 3.8 in Sutton & Barto (2015)), and its degree depends on the environment's constraints on policies. To examine how stochastic sources affect variation in strategic diversity across solutions, we assess how often the policy sets disagree in their action choices. Specifically, we compute the fraction of individual policies' deviations

in action choice relative to the majority-vote ensemble policy of the policy set, averaged over the same 2,500,000 states across all training contexts. Furthermore, to investigate divergence in solution strategies in critical states, we extract the 250,000 most important states, as judged by the action confidence of the highest-performing policy, and examine action deviance within each policy set on this subset.

Studying the divergence in state criticality estimation: One aspect of planning capabilities in sequential decision-making problems is understanding when an action choice is critical to the episode's success and acting accordingly. Therefore, the final experiments aim to measure the divergence in how the policies estimate the importance of choosing the correct action. We extract from each policy set the subset of the 25,000 most important states, as determined by each policy's action confidence. Action confidence is a proxy for criticality estimation because a highly confident policy indicates that its estimates of the episode's remaining steps' success depend heavily on the action choice (Huang et al., 2018). To measure divergence in criticality estimates across the policy set, we compute pairwise intersections between the top 1% most important states for each policy and present the distributions across the policy sets.

4 Experimental Setup

4.1 Policy Training

We train three sets of 100 policies for 200 million environment steps (for each training context and each environment), sharing the same actor-critic architecture, and using a modified implementation of a PPO training scheme[1] in the Pgx version (Koyamada et al., 2023) of the MinAtar testbed (Young & Tian, 2019). The modifications to the script are twofold: first, the addition of a Proximal Feature Optimization (PFO) (Moalla et al., 2024) term to the loss to regularize changes in feature representations of the state, and, second, to allow for the different training contexts, we needed to adjust the seeding strategy for the environment step and reset functions. Adding the PFO term to the loss yielded substantial improvements in policy performance, and therefore, we include it to analyze the divergence of the best set of policies.

The three sets of policies differ solely in how the randomness of the training schemes is fixed. The first context, henceforth referred to as the *baseline* (B), varies both algorithmic randomness (parameter initialization) and environmental randomness (initial-state sampling and transitional randomness) during policy training. In the second context, we isolate the impact of random parameter initialization by fixing the scenarios drawn from the environment and varying the parameter initialization. We refer to the second context as *algorithmic divergence* (AD). The final training context, referred to as *environmental divergence* (ED), isolates environmental randomness by training on the same initial policy parameters across varied scenario samples from the environment.

4.2 The MinAtar Testbed

The MinAtar testbed (Young & Tian, 2019) is a compact version of five Atari games: Asterix, Breakout, Freeway, Seaquest, and Space Invaders. In Asterix, the goal is to collect gold while dodging enemies. In Breakout, the agent's goal is to remove the wall of bricks by deflecting the ball with a paddle. The goal in Freeway is to cross the road as many times as possible within a time limit while dodging cars. In Seaquest, the agent controls a submarine that must rescue divers to replenish its oxygen levels while dodging fish and enemy submarines. In Space Invaders, the agent controls a spaceship under attack by aliens, aiming to shoot down the aliens before they board. We choose the MinAtar testbed because the tasks are varied and the environments vary in complexity: Freeway and Breakout have simpler action spaces, whereas Asterix, Space Invaders, and Seaquest increase in difficulty and complexity, with Seaquest having the largest state-action space. For more information about the environments, see the MinAtar documentation (Young & Tian, 2019).

[1]The PPO implementation is similar to one in`https://github.com/sotetsuk/pgx/blob/87278d2d6e677fd87248c457207b59cfa42e578d/examples/minatar-ppo/train.py`, with an added PFO term in the loss and minor modifications to the random seeding.

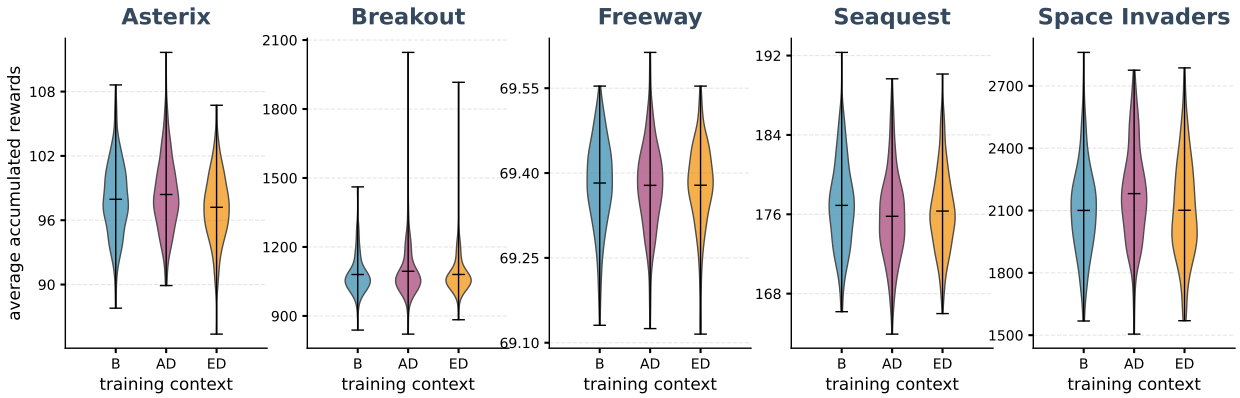

Figure 2: Distributions of average accumulated rewards of the 300 trained policies, for each training context in all environments.

4.3 Technical Setup

The policy training was performed on an NVIDIA DGX-2 with a Tesla V100-SXM3-32GB GPU, while the analyses were conducted on a laptop with an AMD Ryzen 9 7940HS and an NVIDIA GeForce RTX 4070 Laptop GPU, using Python version 3.13.5. To enable experiments at a much larger scale, the experiments were conducted on highly parallelizable versions of the MinAtar testbed (Koyamada et al., 2023) using the JAX framework (Bradbury et al., 2018). Exact implementations of the experiments are available in the supplementary material.

5 Divergence of Trained Policies

This section presents the results of analyses of policy divergence across the three training contexts in the MinAtar environments. First, we present the divergence in policy performance across all 300 trained policies (Section 5.1). Second, two sets of 20 high-performing policies are extracted from each training context, ensuring that the policies within each context are similarly strong (Section 5.2). To investigate the complementarity of these sets, we execute majority-vote ensemble models for each set and compare their performance against the highest-performing policy in that set. Third, the divergence of feature attributions across sets is presented and analyzed (Section 5.3), followed by an investigation of divergence in solution strategies within ensembles (Section 5.4). Finally, we examine how the sets' policies align in their assessments of state criticalities (Section 5.5). We also provide the results of statistical tests comparing all relevant distributions in the appendix (Appendix A).

5.1 Variation in Individual Performances

The distributions of the performances of all the trained 300 policies from the three training contexts are shown in Figure 2. Statistical tests for comparisons between distributions are presented in the appendix, in Table 1. The policies are distributed similarly in terms of mean accumulated rewards across the three training contexts in all environments, with only minor differences in performance distributions; significant (but minor) differences are observed only between the AD training context and the other two in Seaquest and Space Invaders, and between the ED contexts and the other two in Asterix. This indicates that fixing one source of randomness does not substantially change the distributions of the final policies' performances. Consequently, none of the sources of stochasticity seems to contribute more than the other to the final variation of policy performances.

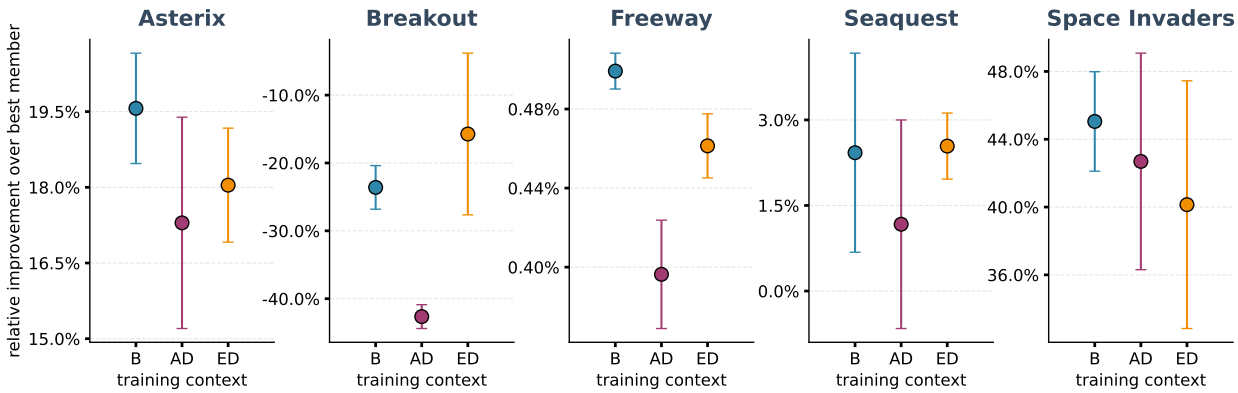

Figure 3: Relative performance increase measured through average accumulated rewards over 1024 episodes over the best policy in the ensemble, with the error bars representing 95% confidence intervals of the standard error of the mean.

5.2 Ensemble Performances

We selected the three sets (obtained by repeating the experiment three times) of 20 highest-performing policies from each training context. The majority-vote ensemble policies of the sets were evaluated over 1024 episodes.

The increases in the average accumulated rewards of all the ensembles relative to their best individual policies are shown in Figure 3. We observe substantial increases in accumulated rewards for the Asterix ($\approx 120\%, 117\%, 118\%$), and Space Invaders ($\approx 145\%, 143\%, 140\%$) environments for the majority-vote ensembles (B, AD, and ED ensembles, respectively). In Seaquest, the performance increases are smaller ($\approx 102\%, 101\%, 103\%$), whereas in Freeway, the performance is comparable to that of the best individual policies. In Freeway, the simplest environment of the testbed, individual policy performance approaches near-optimal for the task, which explains the lack of substantial improvement (a consistent but slight improvement of $\approx 100, 5\%$ across all training contexts) in ensemble performance. However, the majority-vote ensemble policies perform worse than individual policies in Breakout ($77\%, 58\%, 85\%$). In the environments where the ensemble policies outperformed the individuals, we observe that ensemble models from the policy sets trained with varying environment seeding and policy initialization (B) outperform or match the relative performance of ensemble models from the policy sets trained in contexts in which one of these was fixed.

The drop in ensemble performance on Breakout shows that naive majority-vote ensembling can degrade performance. The degrading performance in Breakout is attributable to strategic divergence combined with highly unforgiving environmental dynamics; a single error can often directly lead to termination.

5.3 Alignment in Feature Attributions

To investigate the diversity of state interpretations within the policy sets, we extracted the top five features from the saliencies of the summed action distributions across the individual policies for all 2,500,000 sampled environment states. Statistical tests for comparing distributions are reported in the Appendix, Table 2.

The distributions of average intersections between the policies' top five feature attribution scores over all the states are shown in Figure 4. We observe in Asterix, Breakout, Seaquest, and Space Invaders that sets trained with varying environment seeds and fixed initial model weights (the ED training context) exhibit statistically significantly higher alignment in their top five features, with a non-negligible effect size. In Freeway, the alignment is similar across the training contexts. Across all environments except Freeway, the distributions of feature attributions for policy predictions differ significantly for policies trained with different parameter initializations (B and AD), but the effect size is substantially smaller. We also observe that the overlap is generally quite low across the board ($\approx 35\%, 20\%, 40\%, 20\%, 30\%$, in Asterix, Breakout,

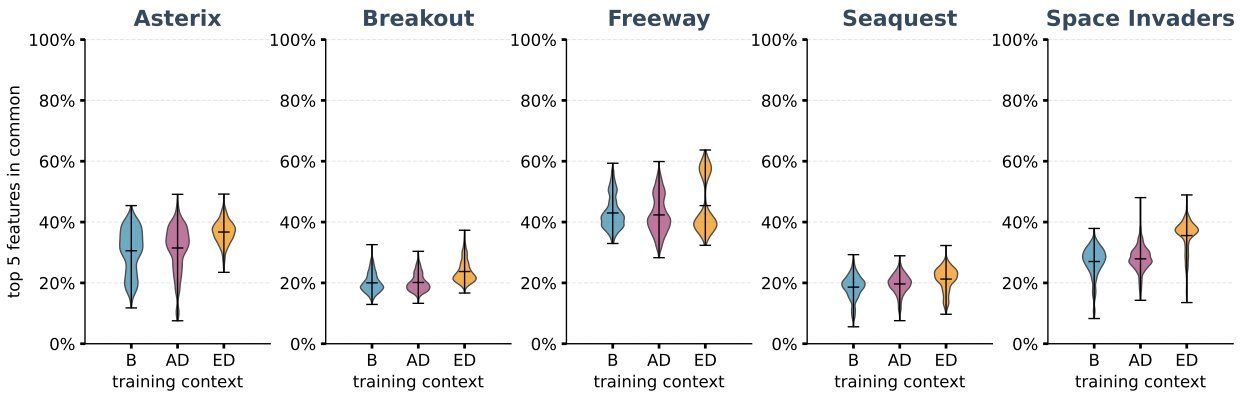

Figure 4: Distributions of average pairwise intersections of the top five features in the ensembles of policies.

Freeway, Seaquest, and Space Invaders, respectively), showing that the policies highlight different features in the saliency maps.

The greater alignment in feature attributions across ED policies suggests that random parameter initialization leads to greater misalignment than varying the sampled scenarios. Thus, the results suggest that the algorithmic divergence has a greater impact on the level of misalignment in feature attributions observed in similarly trained high-performing policies, although they also indicate large divergence in feature attributions across all training contexts.

5.4 Disaccord in Action Choice

To investigate the diversity of solution strategies within the policy sets, we measure disagreement in action choices among the policies' individual action choices across the 2,500,000 sampled environment states. Specifically, we compute the fraction of deviance in action choices between individual policies and the majority-vote ensemble policy across all states and the most critical states. In states deemed critical, the choice of action is considered highly important to the episode's success, making disagreement over actions in such states particularly revealing of differing interpretations of the environment between the policies. Again, we report statistical tests comparing distributions in the Appendix, Tables 3 and 4.

We show the distributions of the fraction of action deviance from the majority-vote ensemble over all states in the top row of Figure 5. The bottom row shows the action deviance for the subset of the 10% most critical states, as judged by the strongest policy in the set. We observe similar levels of disagreement within the three policy sets over all 2,500,000 states ($\approx 21\%, 27\%, 5\%, 23\%, 22\%$ for Asterix, Breakout, Freeway, Seaquest, and Space Invaders, respectively). For the subset of the 250,000 most critical states, we observe less disagreement in action choice (averages are $\approx 2\%, 8\%, 0.1\%, 12\%, 5\%$ for Asterix, Breakout, Freeway, Sequest, and Space Invaders, respectively). Across training contexts, we observe similar distributions in Asterix, Freeway, Seaquest, and Space Invaders, with only marginal differences between the policy sets. In Breakout, we observe a minor increase in the action disagreement between ED policies compared to the other two training contexts.

The similarities in action disagreements across all policy sets indicate that neither the initial agent nor the specific realizations of the environment are the causes of the observed diversity in policy solutions. Thus, the *interaction between the agent and the environment* during training must be highly sensitive to the state of the learning system, both the current network weights and the order and content of environment states exposed to it, thereby yielding diverse experiences and learning outcomes. Also, the degrees of action disaccord within the sets demonstrate the existence and prevalence of distinct, high-performing solution strategies in the environments.

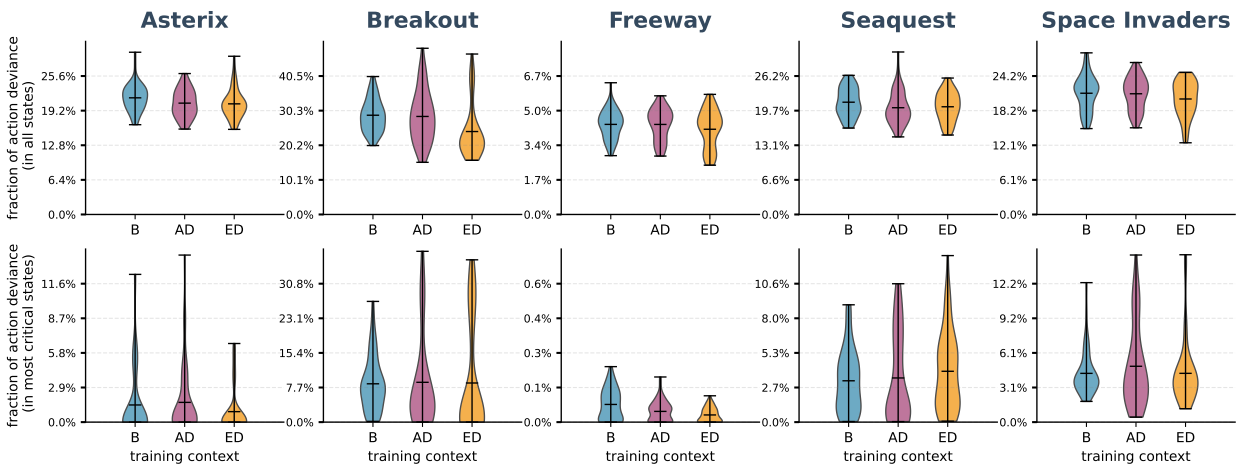

Figure 5: Fractions of action deviance from the majority vote ensemble policy of each policy in the set for all 2,500,000 states in the trajectory (top row), and the 250,000 most important states as deemed by the highest-performing policy in the set (bottom row).

5.5 Alignment in State Criticality

The criticality of a state is the degree to which the choice of action in that state is critical to the success of the episode. One measure of the state's criticality is the policy's confidence in the chosen action, quantified by the entropy of its action distribution. To measure how the ensembles align in their assessments of state criticality, we extract the 1% most critical states for each policy in each ensemble and calculate the intersection of these sets for each pair of policies. As before, we report the results of our statistical significance tests in the Appendix, Table 5.

The distributions of pairwise intersections in the top 1% most critical states (2,500 states) between the policies of each policy set, for all the environments, are shown in Figure 6. Note that the ED policies exhibit marginally higher alignment in the most critical states across most environments; only in Seaquest do the AD policies align more. While our statistical analyses indicate that these are statistically distinct (Appendix, Table 5), the pairwise overlaps remain highly similar in overall magnitude across all training contexts.

Although we observe a consistent but minor increase in alignment for the ED policies, the results suggest that the three contexts converge to the same overlap distributions for state-criticality estimates. Again, as with performance distributions and action agreement, we observe that the three contexts are similarly aligned in their state-criticality estimates, further suggesting that neither the initial agent nor the environment realizations affects the distributional outcome of policies, but rather shows that the dominant source of divergence in state estimations lies within the agent-environment interaction.

6 Discussion

On the performances of the ensembles: We found that the majority-vote ensembles of the policy sets improved performance compared to the best individual policy in Asterix, Seaquest, and Space Invaders, whereas ensembling in Breakout degraded performance. Although there were minor performance differences across training contexts, they were inconsistent and small. In Breakout, we hypothesize that the performance degradation is linked to high disagreement in action choice in high-importance states (Figure 5, second row). Because Breakout requires highly precise action sequences to avoid immediate failure, a simple majority vote among strategically divergent policies dilutes the precise execution strategy of individual policies. We observe even higher action disaccord among the Seaquest policies, which can be attributed to its larger action space and the availability of multiple viable strategies. Unlike in Breakout, where a single suboptimal

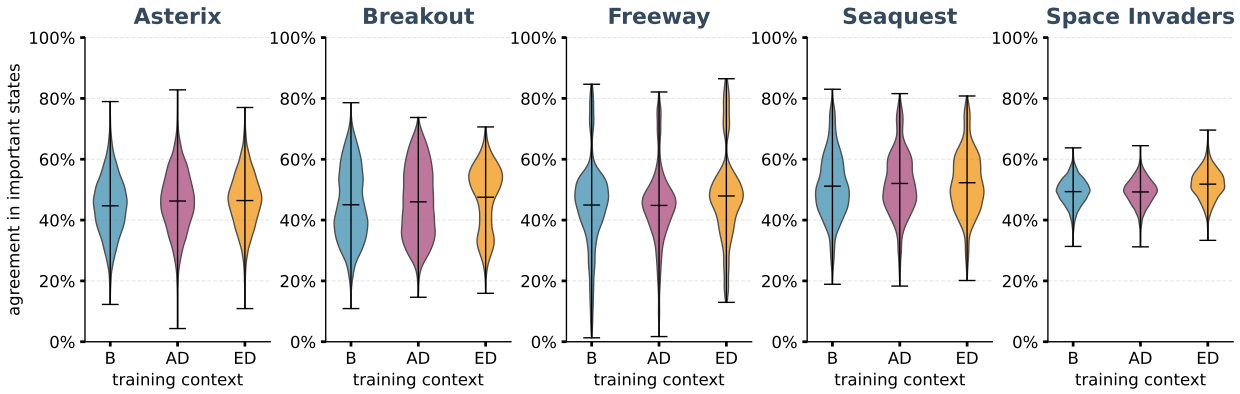

Figure 6: Distributions of pairwise intersections of the 1% highest importance states as deemed by each policy in the ensembles.

action is often fatal, in Seaquest, the agents can often recover from errors, making strategic divergence less detrimental to overall performance. This highlights a limitation to naive ensembling in RL; when there is high strategic divergence, ensembling can be actively detrimental.

On the alignment in feature attributions: The set of policies derived from the ED training context consistently had a larger alignment in their feature attributions compared to the other training contexts. In contrast, the alignments were similar between policy sets derived from the AD and B regimes (see Figure 4). We interpret these results as evidence that the parameter initialization is a stronger determinant of the final feature representation than the sampled scenarios, possibly suggesting that the initial parameterization may render some potentially useful features inaccessible. However, given the low dimensionality of the MinAtar observation space (a binary $C \times 10 \times 10$ where $C \leq 10$), an alternative hypothesis is that the ED policies are biased by identical initialization toward memorizing exact spatial coordinates (pixels) rather than extracting generalized features. Thus, weight initialization may not only restrict the accessible features but also actively anchor spatial representations in low-dimensional settings.

On disagreement in action choice within policy sets: The distributions of action deviance within the sets are similar across training contexts (see Figure 5), both on the full set of 2,500,000 states and in the 10% most important states. As expected, we observe less action deviance in the 10% most important states than in the full set of states across all environments, showing that the policies converge on the same actions when stakes are higher. Despite greater alignment in feature attributions, we observe that high action disaccord persists in the ED context. This suggests a decoupling between state representation and strategic execution: although the algorithmic initialization biases the agents toward using more of the same features, their operational strategies still diverge to the same extent as in other training contexts due to compounding intra-training stochasticity (e.g., action sampling and mini-batch selection). Overall, these results demonstrate that fixing initialization and environment sampling is insufficient to constrain policy behavior, highlighting that the PPO algorithm steers policies toward a multiplicity of distinct, yet viable, strategic optima.

On the divergence of state criticality assessments: The distributions of overlaps in the 1% most important states (Figure 6) show only marginal differences between training contexts. This indicates that controlling for the initial untrained agent or for environmental sampling fails to unify the way policies evaluate risk and state importance. Even when policies share identical initial weights or train on identical environment realizations, their assessments of state criticality diverge to the same extent as those of fully uncontrolled baselines. This reinforces the conclusion that the state importance measure is highly subjective, heavily dependent on the specifics of the stochastically driven training trajectory. Furthermore, this subjectivity is linked to the observed disaccord in the action distributions; as agents develop divergent operational strategies, their underlying value estimates and perceptions of criticality will also diverge. Ultimately, this

highlights that such path dependence is not merely an artifact of a specific random seed but rather an inherent characteristic arising from the highly sensitive, recursive feedback loop between stochastic exploration and policy updates in the PPO algorithmic framework.

On limitations and future work: The complexity of online RL training algorithms introduces potential concerns about the validity and generalizability of the results.

First, to mitigate potential artifacts arising from the specific seed used for the environment and for parameter initialization in the AD and ED contexts, respectively, we trained three sets of 100 policies in each training context. Second, our experiments were conducted on the MinAtar testbed, and our results may not generalize to other environments, especially environments with continuous state and/or action spaces. However, the environments of the MinAtar testbed provide RL tasks of varying complexity and difficulty and are highly parallelizable, making them suitable for studies such as this. Third, we chose to limit ourselves to PPO policies, one to reduce computational effort (we trained and evaluated 4,500 policies in total in this project) and two to limit the number of experimental variables. Adding more Deep RL approaches, such as DQN (Mnih et al., 2015) or SAC (Haarnoja et al., 2018), would provide a broader context for interpreting the result and could be explored in future work. Fourth, for this work, we examine a single policy architecture across all training contexts and environments. Because the algorithmic divergence depends on the policy architecture's functional space, the analysis results will vary with the chosen architecture. We chose a well-performing policy architecture that is consistent across all environments to enable comparisons of results between environments. Furthermore, we suggest that investigations into architectural dependence on divergence of solutions are more practically conducted in the context of supervised learning. Finally, extracting and examining policies across multiple stages of the training process to analyze the evolution of policy sets is an interesting avenue for future work.

7 Related Work

Our work is related to three active research areas. The Rashomon effect of supervised learning (SL) (Breiman, 2001) demonstrates the multiplicity of machine learning (ML) models for identical learning problems. Because ML models are also used in Deep RL, the Rashomon effect in SL provides a strong foundation for extending this work to Deep RL. Furthermore, the analysis of policy divergence draws on concepts from XRL, such as critical states (Spielberg & Azaria, 2019), and is intended to contribute to the body of knowledge. Finally, research on the instability of RL training algorithms is closely related to our work. However, the field is more focused on failed training trials and reproducibility, whereas we investigate the divergence of trained policies.

7.1 The Rashomon Effect

The notion that statistical learners can extract high-performing, varied interpretations from the same data was coined the *Rashomon effect* by Breiman (2001). Müller et al. (2023) investigate the Rashomon effect in the context of explainable machine learning; specifically, they provide empirical evidence for the divergence of models on fixed datasets, and thus provide evidence of the lack of transfer value of explanations between models. Laberge et al. (2023) identify essential features based on consensus within the Rashomon set.

Our work is adjacent to investigations of the Rashomon effect in supervised learning and its impact on explainable ML, because *the multiplicity of high-performing models* is inherited in the XRL context. Because of the additional temporal dimension in sequential decision-making tasks, the policies' freedom to learn different solution strategies and to estimate state and action values may amplify their impact in the RL context relative to the SL context.

7.2 XRL/Critical States & State Importance

A critical state is one in which the choice of action strongly affects the episode's success (Spielberg & Azaria, 2019), one in which the policy must behave reliably. Because sequential decision-making tasks are often complex and require extensive planning to understand the long-term consequences of decisions, identifying

critical states is not always straightforward. Still, for developers and researchers seeking to validate or understand policy behavior, these states constitute a fundamental form of task awareness. Therefore, methods for identifying these states are pivotal to XRL (Gross & Spieker, 2025).

RL policies are fundamentally derived from estimates of the *value* of actions and states. This intrinsic value estimation allows the extraction of the *policies' estimate of the state's criticality*. The *state importance* measures exactly that, and is often derived from the spread of the value of different actions from the same state. For example, in Amir & Amir (2018), the state importance is defined as the span between the maximum action-value and minimum action-value in a particular state. However, under PPO policies, the action-value function is less accessible, so we follow Huang et al. (2018) and instead use the entropy of the action distributions. Both works aim to establish trust and explain policy behavior by focusing on the policies' assessments of state criticality.

7.3 Instability of Training Schemes in Reinforcement Learning

The instability of training schemes in RL is a well-known problem in reproducible RL research and a source of trust issues for practitioners (Clary et al., 2019; Chan et al., 2019). In Henderson et al. (2018), they provide empirical evidence of a reproducibility issue, investigate statistical measures to provide a more reliable assessment of procedures, and, most relevant to our work, show considerable variability in performance due to the inherent stochasticity of training schemes. Building on this work, Lynnerup et al. (2020) present a practical assessment of the reproducibility and replicability of RL on a continuous control task. Also, there are works on *why* training algorithms sometimes fail to learn a good policy. Moalla et al. (2024) investigate why PPO policies suffer from *capacity loss*, a weakened ability to learn from new data, in long training runs. Their analysis suggests that a rank collapse in learned state representations hinders continual learning because the forward pass loses information from the observations. They propose a method that employs a trust region in the latent feature space to mitigate loss of learning capacity. Closely related is Bjorck et al.'s work (Bjorck et al., 2021), which shows that normalizing feature representations reduces the frequency of failed training runs, i.e., training runs in which the policies fail to learn a reasonably well-performing policy. Perhaps most closely related to our work is (Jang et al., 2025), which investigates the performance variability of PPO agents under the hypothesis that it is caused by diverging experience and reports a correlation between diverging experience and performance variability.

Similarly, our work aims to provide further insight into RL policy training by investigating how sources of stochasticity affect the policies' performance variability. Furthermore, it shifts the investigation to how these sources affect the divergence in how the policies interpret and plan in their environments.

8 Conclusions

In this work, we investigate the sources of policy divergence in RL policies by isolating the effects of environmental scenario sampling and parameter initialization. By training and evaluating sets of PPO policies across the MinAtar testbed, we demonstrated that controlling these initial conditions does not constrain the divergence of final policies. Although fixing the initial weights led to more aligned policies in their feature attributions, they still exhibited severe disagreement in their action choices and subjective assessments of state criticality, suggesting a decoupling between state representation and strategic execution. Our results indicate that the divergence among similarly trained PPO policies is not primarily driven by the initial random seed or environmental configuration, but rather by the high sensitivity of the recursive feedback loop between stochastic exploration and policy updates within the PPO algorithm. This inherent sensitivity of PPO steers policies toward a multiplicity of distinct yet viable strategic optima, many of which are suboptimal. Such algorithmic unpredictability challenges the community's reliance on individual executions for benchmarking algorithmic innovations, highlighting the importance of large-scale evaluation protocols.

In conclusion, our investigation shows that a single PPO implementation yields a vast landscape of possible policies.

References

Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron C Courville, and Marc Bellemare. Deep Reinforcement Learning at the Edge of the Statistical Precipice. In *Advances in Neural Information Processing Systems*, volume 34, pp. 29304–29320. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/hash/f514cec81cb148559cf475e7426eed5e-Abstract.html.

Dan Amir and Ofra Amir. HIGHLIGHTS: Summarizing Agent Behavior to People. July 2018. URL https://www.semanticscholar.org/paper/HIGHLIGHTS%3A-Summarizing-Agent-Behavior-to-People-Amir-Amir/f611acc0ee253f5903dafe30bbcd65f823374448.

Johan Bjorck, Carla P. Gomes, and Kilian Q. Weinberger. Is High Variance Unavoidable in RL? A Case Study in Continuous Control. *ArXiv*, October 2021. URL https://www.semanticscholar.org/paper/Is-High-Variance-Unavoidable-in-RL-A-Case-Study-in-Bjorck-Gomes/c578a295ba3456bebc651f1f516e34ea6e7a5fe3.

James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal MacLaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs, 2018. URL http://github.com/jax-ml/jax.

Leo Breiman. Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). *Statistical Science*, 16(3):199–231, August 2001. ISSN 0883-4237, 2168-8745. doi: 10.1214/ss/1009213726. URL https://projecteuclid.org/journals/statistical-science/volume-16/issue-3/Statistical-Modeling--The-Two-Cultures-with-comments-and-a/10.1214/ss/1009213726.full. Publisher: Institute of Mathematical Statistics.

Stephanie C. Y. Chan, Samuel Fishman, Anoop Korattikara, John Canny, and Sergio Guadarrama. Measuring the Reliability of Reinforcement Learning Algorithms. September 2019. URL https://openreview.net/forum?id=SJlpYJBKvH.

Kaleigh Clary, Emma Tosch, John Foley, and David D. Jensen. Let's Play Again: Variability of Deep Reinforcement Learning Agents in Atari Environments. *ArXiv*, April 2019. URL https://www.semanticscholar.org/paper/Let%27s-Play-Again%3A-Variability-of-Deep-Reinforcement-Clary-Tosch/57243359b2b6794835eb4d4be3f67b6da5f7e2a1.

Logan Engstrom, Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Firdaus Janoos, Larry Rudolph, and Aleksander Madry. Implementation Matters in Deep RL: A Case Study on PPO and TRPO. April 2020. URL https://iclr.cc/virtual_2020/poster_r1etN1rtPB.html.

Dennis Gross and Helge Spieker. PCTL model checking for temporal RL policy safety explanations. In *SAC*, pp. 1514–1521. ACM, 2025.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 1861–1870. PMLR, July 2018. URL https://proceedings.mlr.press/v80/haarnoja18b.html. ISSN: 2640-3498.

Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. Deep reinforcement learning that matters. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18, pp. 3207–3214, New Orleans, Louisiana, USA, February 2018. AAAI Press. ISBN 978-1-57735-800-8.

Sandy H. Huang, Kush Bhatia, Pieter Abbeel, and Anca D. Dragan. Establishing Appropriate Trust via Critical States, October 2018. URL http://arxiv.org/abs/1810.08174. arXiv:1810.08174 [cs].

Riashat Islam, Peter Henderson, Maziar Gomrokchi, and Doina Precup. Reproducibility of Benchmarked Deep Reinforcement Learning Tasks for Continuous Control. *ArXiv*, June 2017. URL https://www.semanticscholar.org/paper/Reproducibility-of-Benchmarked-Deep-Reinforcement-Islam-Henderson/dcc2941b5d7a70acf8bd609570e4b6f106ecabc4.

Sooyoung Jang, Hyung-Il Kim, Sooyoung Jang, and Hyung-Il Kim. Entropy-Aware Model Initialization for Effective Exploration in Deep Reinforcement Learning. *Sensors*, 22(15), August 2022. ISSN 1424-8220. doi: 10.3390/s22155845. URL https://www.mdpi.com/1424-8220/22/15/5845.

Sooyoung Jang, Ahyun Lee, Sooyoung Jang, and Ahyun Lee. Unpacking Performance Variability in Deep Reinforcement Learning: The Role of Observation Space Divergence. *Applied Sciences*, 15(15), July 2025. ISSN 2076-3417. doi: 10.3390/app15158247. URL https://www.mdpi.com/2076-3417/15/15/8247. Company: Multidisciplinary Digital Publishing Institute Distributor: Multidisciplinary Digital Publishing Institute Institution: Multidisciplinary Digital Publishing Institute Label: Multidisciplinary Digital Publishing Institute Publisher: publisher.

Sotetsu Koyamada, Shinri Okano, Soichiro Nishimori, Yu Murata, Keigo Habara, Haruka Kita, and Shin Ishii. Pgx: Hardware-Accelerated Parallel Game Simulators for Reinforcement Learning. *Advances in Neural Information Processing Systems*, 36:45716–45743, December 2023. URL https://papers.nips.cc/paper_files/paper/2023/hash/8f153093758af93861a74a1305dfdc18-Abstract-Datasets_and_Benchmarks.html.

Gabriel Laberge, Yann Pequignot, Alexandre Mathieu, Foutse Khomh, and Mario Marchand. Partial Order in Chaos: Consensus on Feature Attributions in the Rashomon Set. *Journal of Machine Learning Research*, 24(364):1–50, 2023. ISSN 1533-7928. URL http://jmlr.org/papers/v24/23-0149.html.

Nicolai A. Lynnerup, Laura Nolling, Rasmus Hasle, and John Hallam. A Survey on Reproducibility by Evaluating Deep Reinforcement Learning Algorithms on Real-World Robots. In *Proceedings of the Conference on Robot Learning*, pp. 466–489. PMLR, May 2020. URL https://proceedings.mlr.press/v100/lynnerup20a.html. ISSN: 2640-3498.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level Control through Deep Reinforcement Learning. *Nature*, 518(7540):529–533, February 2015. ISSN 1476-4687. doi: 10.1038/nature14236. URL https://www.nature.com/articles/nature14236. Publisher: Nature Publishing Group.

Skander Moalla, Andrea Miele, Daniil Pyatko, Razvan Pascanu, and Caglar Gulcehre. No Representation, No Trust: Connecting Representation, Collapse, and Trust Issues in PPO. *Advances in Neural Information Processing Systems*, 37:69652–69699, December 2024. doi: 10.52202/079017-2226. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/81166fbd9cc5adf14031cdb69d3fd6a8-Abstract-Conference.html.

Miguel Morales. *Grokking Deep Reinforcement Learning*, volume 1. Manning Publications Co., New York, first edition edition, 2020. ISBN 978-1-61729-545-4.

Sebastian Müller, Vanessa Toborek, Katharina Beckh, Matthias Jakobs, Christian Bauckhage, and Pascal Welke. An Empirical Evaluation of the Rashomon Effect in Explainable Machine Learning, June 2023. URL http://arxiv.org/abs/2306.15786. arXiv:2306.15786 [cs].

Marcel Panzer and Benedict Bender. Deep Reinforcement Learning in Production Systems: a Systematic Literature Review. *International Journal of Production Research*, 60(13):4316–4341, July 2022. ISSN 0020-7543. doi: 10.1080/00207543.2021.1973138. URL https://doi.org/10.1080/00207543.2021.1973138. _eprint: https://doi.org/10.1080/00207543.2021.1973138.

Marco Antonio Paz Ramos and Axel Busboom. Systematic Review of Reinforcement Learning in Process Industries: A Contextual and Taxonomic Approach. *Applied Sciences*, 15(24), December 2025. ISSN 2076-3417. doi: 10.3390/app152412904. URL https://www.mdpi.com/2076-3417/15/24/12904.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal Policy Optimization Algorithms, August 2017. URL http://arxiv.org/abs/1707.06347. arXiv:1707.06347 [cs].

David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. Mastering the Game of Go with Deep Neural Networks and Tree Search. *Nature*, 529(7587):484–489, January 2016. ISSN 1476-4687. doi: 10.1038/nature16961.

Yitzhak Spielberg and Amos Azaria. The Concept of Criticality in Reinforcement Learning. In *2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI)*, pp. 251–258, November 2019. doi: 10.1109/ICTAI.2019.00043. URL https://ieeexplore.ieee.org/document/8995362. ISSN: 2375-0197.

Richard S Sutton and Andrew G Barto. *Reinforcement Learning: An Introduction*. The MIT Press, Cambridge, Massachusetts, second edition edition, 2015.

Emanuel Todorov, Tom Erez, and Yuval Tassa. MuJoCo: A physics engine for model-based control. *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 5026–5033, October 2012. doi: 10.1109/IROS.2012.6386109. URL http://ieeexplore.ieee.org/document/6386109/. Conference Name: 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2012) ISBN: 9781467317368 9781467317375 9781467317351 Place: Vilamoura-Algarve, Portugal Publisher: IEEE.

M.A. Wiering and H. Van Hasselt. Ensemble Algorithms in Reinforcement Learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 38(4):930–936, August 2008. ISSN 1083-4419. doi: 10.1109/TSMCB.2008.920231. URL http://ieeexplore.ieee.org/document/4509588/.

Danny Wood, Tingting Mu, Andrew M. Webb, Henry W. J. Reeve, Mikel Luján, and Gavin Brown. A Unified Theory of Diversity in Ensemble Learning. *J. Mach. Learn. Res.*, 24(1):359:17302–359:17350, January 2023. ISSN 1532-4435.

Kenny Young and Tian Tian. MinAtar: An Atari-Inspired Testbed for Thorough and Reproducible Reinforcement Learning Experiments, June 2019. URL http://arxiv.org/abs/1903.03176. arXiv:1903.03176 [cs].

A Appendix

The appendix reports statistical tests for pairwise comparisons between the distributions shown in Section 5.

A.1 Tests of Significance of Results

In this section, we present statistical significance tests for pairwise comparisons of results across each training setup. We perform a Mann-Whitney U test to assess whether the underlying populations are stochastically equal, and report Cliff's delta to quantify the extent of overlap between the distributions.

First, we present the results of our statistical analysis of pairwise performance comparisons between training setups in Table 1, which correspond to the results shown in Figure 2.

Second, Table 2 presents the statistical analysis of the pairwise comparisons between the training setups, based on the distributions of average pairwise intersections of the top five features for all possible pairings of two policies within the policy sets visualized in Figure 6.

Third, the statistical tests on the results from the top and bottom rows of Figure 5 are tabulated in Table 3 and Table 4, respectively. They are tabulating the statistical tests between the training setups' distributions

Table 1: Statistical tests on the pairwise comparisons between the performances of each training setup's 300 policies. Distributions of performances are shown in Figure 2.

Environment	Comparison	Test	Statistic	p-value	Significant	Cliff's δ
Asterix	B vs ED	Mann-Whitney U	50412.00	0.011	Yes	0.120 (negligible)
	B vs AD	Mann-Whitney U	42116.00	0.174	No	-0.064 (negligible)
	ED vs AD	Mann-Whitney U	36924.00	< 0.001	Yes	-0.179 (small)
Breakout	B vs ED	Mann-Whitney U	43940.00	0.826	No	-0.010 (negligible)
	B vs AD	Mann-Whitney U	41944.00	0.302	No	-0.049 (negligible)
	ED vs AD	Mann-Whitney U	42175.00	0.433	No	-0.037 (negligible)
Freeway	B vs ED	Mann-Whitney U	46290.50	0.543	No	0.029 (negligible)
	B vs AD	Mann-Whitney U	47247.00	0.290	No	0.050 (negligible)
	ED vs AD	Mann-Whitney U	46139.00	0.592	No	0.025 (negligible)
Seaquest	B vs ED	Mann-Whitney U	48110.00	0.078	No	0.084 (negligible)
	B vs AD	Mann-Whitney U	46322.00	0.002	Yes	0.147 (small)
	ED vs AD	Mann-Whitney U	43343.00	0.130	No	0.073 (negligible)
Space Invaders	B vs ED	Mann-Whitney U	45051.00	0.868	No	0.008 (negligible)
	B vs AD	Mann-Whitney U	35811.00	< 0.001	Yes	-0.196 (small)
	ED vs AD	Mann-Whitney U	36069.00	< 0.001	Yes	-0.190 (small)

Table 2: Statistical tests on the comparisons between the distributions of pairwise overlaps in the top five features for all pairings within the policy sets from the different training setups.

Environment	Comparison	Test	Statistic	p-value	Significant	Cliff's δ
Asterix	B vs ED	Mann-Whitney U	8205308.00	< 0.001	Yes	-0.495 (large)
	B vs AD	Mann-Whitney U	14942202.00	< 0.001	Yes	-0.080 (negligible)
	ED vs AD	Mann-Whitney U	23299234.00	< 0.001	Yes	0.434 (medium)
Breakout	B vs ED	Mann-Whitney U	6654506.00	< 0.001	Yes	-0.590 (large)
	B vs AD	Mann-Whitney U	15703996.00	0.002	Yes	-0.033 (negligible)
	ED vs AD	Mann-Whitney U	25802462.00	< 0.001	Yes	0.588 (large)
Freeway	B vs ED	Mann-Whitney U	15447628.00	< 0.001	Yes	-0.049 (negligible)
	B vs AD	Mann-Whitney U	17416246.00	< 0.001	Yes	0.072 (negligible)
	ED vs AD	Mann-Whitney U	18494222.00	< 0.001	Yes	0.138 (negligible)
Seaquest	B vs ED	Mann-Whitney U	9632262.00	< 0.001	Yes	-0.407 (medium)
	B vs AD	Mann-Whitney U	13714682.00	< 0.001	Yes	-0.156 (small)
	ED vs AD	Mann-Whitney U	21178022.00	< 0.001	Yes	0.304 (small)
Space Invaders	B vs ED	Mann-Whitney U	3464106.00	< 0.001	Yes	-0.787 (large)
	B vs AD	Mann-Whitney U	15475754.00	< 0.001	Yes	-0.047 (negligible)
	ED vs AD	Mann-Whitney U	28538574.00	< 0.001	Yes	0.757 (large)

of the fraction of states that each policy in the set of policies deviated from the majority-vote ensemble policy across all states (Table 3) and in the subset of the 250,000 states deemed of highest importance (Table 4).

Table 3: Statistical tests on the comparisons between action-deviance distributions for all states.

Environment	Comparison	Test	Statistic	p-value	Significant	Cliff's δ
Asterix	B vs ED (All states)	Mann-Whitney U	206284.00	< 0.001	Yes	0.146 (negligible)
	B vs AD (All states)	Mann-Whitney U	207345.50	< 0.001	Yes	0.152 (small)
	ED vs AD (All states)	Mann-Whitney U	180315.50	0.958	No	0.002 (negligible)
Breakout	B vs ED (All states)	Mann-Whitney U	216192.00	< 0.001	Yes	0.201 (small)
	B vs AD (All states)	Mann-Whitney U	184802.50	0.424	No	0.027 (negligible)
	ED vs AD (All states)	Mann-Whitney U	152546.50	< 0.001	Yes	-0.153 (small)
Freeway	B vs ED (All states)	Mann-Whitney U	203528.50	< 0.001	Yes	0.131 (negligible)
	B vs AD (All states)	Mann-Whitney U	172525.50	0.213	No	-0.042 (negligible)
	ED vs AD (All states)	Mann-Whitney U	150953.50	< 0.001	Yes	-0.161 (small)
Seaquest	B vs ED (All states)	Mann-Whitney U	212969.00	< 0.001	Yes	0.183 (small)
	B vs AD (All states)	Mann-Whitney U	217762.50	< 0.001	Yes	0.210 (small)
	ED vs AD (All states)	Mann-Whitney U	186441.50	0.283	No	0.036 (negligible)
Space Invaders	B vs ED (All states)	Mann-Whitney U	205142.00	< 0.001	Yes	0.140 (negligible)
	B vs AD (All states)	Mann-Whitney U	178416.00	0.792	No	-0.009 (negligible)
	ED vs AD (All states)	Mann-Whitney U	151525.00	< 0.001	Yes	-0.158 (small)

Table 4: Statistical tests on the comparisons between action-deviance distributions for the 10% most important states.

Environment	Comparison	Test	Statistic	p-value	Significant	Cliff's δ
Asterix	B vs ED (Most critical states)	Mann-Whitney U	165522.00	0.003	Yes	-0.080 (negligible)
	B vs AD (Most critical states)	Mann-Whitney U	137873.50	< 0.001	Yes	-0.234 (small)
	ED vs AD (Most critical states)	Mann-Whitney U	149374.00	< 0.001	Yes	-0.170 (small)
Breakout	B vs ED (Most critical states)	Mann-Whitney U	120946.00	< 0.001	Yes	-0.328 (small)
	B vs AD (Most critical states)	Mann-Whitney U	146938.00	< 0.001	Yes	-0.184 (small)
	ED vs AD (Most critical states)	Mann-Whitney U	207105.00	< 0.001	Yes	0.151 (small)
Freeway	B vs ED (Most critical states)	Mann-Whitney U	213280.00	< 0.001	Yes	0.185 (small)
	B vs AD (Most critical states)	Mann-Whitney U	202555.50	< 0.001	Yes	0.125 (negligible)
	ED vs AD (Most critical states)	Mann-Whitney U	168815.00	0.002	Yes	-0.062 (negligible)
Seaquest	B vs ED (Most critical states)	Mann-Whitney U	176992.00	0.614	No	-0.017 (negligible)
	B vs AD (Most critical states)	Mann-Whitney U	165599.50	0.016	Yes	-0.080 (negligible)
	ED vs AD (Most critical states)	Mann-Whitney U	168904.00	0.060	No	-0.062 (negligible)
Space Invaders	B vs ED (Most critical states)	Mann-Whitney U	191609.00	0.053	No	0.064 (negligible)
	B vs AD (Most critical states)	Mann-Whitney U	163930.00	0.007	Yes	-0.089 (negligible)
	ED vs AD (Most critical states)	Mann-Whitney U	149028.00	< 0.001	Yes	-0.172 (small)

Finally, in Table 5, we have tabulated the statistical tests on the comparisons between the training setups' ensembles distributions of pairwise overlaps in the 1% most critical states as deemed by each policy in the sets, corresponding to the results from Figure 6.

A.2 Implementation of Training Scheme

The implementation of the PPO learning algorithm is in the supplementary material. In this section, we provide the hyperparameters used across environments and training contexts in Table 6. To ensure clear communication, we first remind ourselves of the training objective function to maximize.

Let π_θ be the policy, V_θ the value function, and ϕ_θ be the penultimate actor features. The rollout data uses the parameters at the collection time, θ_{old}, and they are stored as $\log \pi_{\theta_{\text{old}}}(a_t|s_t)$, $V_{\theta_{\text{old}}}(s_t)$, $\phi_{\theta_{\text{old}}}(s_t)$. The estimate for the temporal difference is given as

$$\delta_t = r_t + \gamma V_{\theta_{\text{old}}}(s_{t+1})(1 - 1_{\text{done},t}) - V_{\theta_{\text{old}}}(s_t),$$

which leads to the Generalized Advantage Estimate (GAE) being

$$\hat{A}_t = \delta_t + \gamma\lambda(1 - 1_{\text{done}})\hat{A}_{t+1},$$

Table 5: Statistical tests on the comparisons between the distributions of pairwise overlaps in the top 1% important states for all pairings within the policy sets from the different training setups.

Environment	Comparison	Test	Statistic	p-value	Significant	Cliff's δ
Asterix	B vs ED	Mann-Whitney U	14544820.00	< 0.001	Yes	-0.105 (negligible)
	B vs AD	Mann-Whitney U	14763944.00	< 0.001	Yes	-0.091 (negligible)
	ED vs AD	Mann-Whitney U	16394242.00	0.396	No	0.009 (negligible)
Breakout	B vs ED	Mann-Whitney U	14101946.00	< 0.001	Yes	-0.132 (negligible)
	B vs AD	Mann-Whitney U	15499766.00	< 0.001	Yes	-0.046 (negligible)
	ED vs AD	Mann-Whitney U	17622154.00	< 0.001	Yes	0.085 (negligible)
Freeway	B vs ED	Mann-Whitney U	14846336.00	< 0.001	Yes	-0.086 (negligible)
	B vs AD	Mann-Whitney U	17113434.00	< 0.001	Yes	0.053 (negligible)
	ED vs AD	Mann-Whitney U	18578926.00	< 0.001	Yes	0.144 (negligible)
Seaquest	B vs ED	Mann-Whitney U	15140018.00	< 0.001	Yes	-0.068 (negligible)
	B vs AD	Mann-Whitney U	15400380.00	< 0.001	Yes	-0.052 (negligible)
	ED vs AD	Mann-Whitney U	16524686.00	0.111	No	0.017 (negligible)
Space Invaders	B vs ED	Mann-Whitney U	11464778.00	< 0.001	Yes	-0.294 (small)
	B vs AD	Mann-Whitney U	16487422.00	0.168	No	0.015 (negligible)
	ED vs AD	Mann-Whitney U	21284010.00	< 0.001	Yes	0.310 (small)

which are normalized as

$$\tilde{A}_t = \frac{\hat{A}_t - \hat{E}\Big[\hat{A}_t\Big]}{\sqrt{Var(\hat{A}_t)} + 10^{-8}}.$$

Value targets are calculated as

$$\hat{R}_t = \hat{A}_t + V_{\theta_{\text{old}}}(s_t),$$

and the importance ratio is

$$r_t(\theta) = \exp(\log \pi_\theta(a_t|s_t) - \log \pi_{\theta_{\text{old}}}(a_t|s_t)).$$

The first component of the loss, the *clipped actor loss* (with the clipping parameter ϵ), is

$$L^{\text{CLIP}}(\theta) = -\hat{\mathbb{E}}_t\big[\min(r_t(\theta)\tilde{A}_t, \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\tilde{A}_t\big].$$

The second component, the *clipped value loss*, which shares the clipping parameter ϵ, makes use of clipped value estimates given as

$$V_\theta^{\text{clip}}(s_t) = V_{\theta_{\text{old}}}(s_t) + \text{clip}(V_\theta(s_t) - V_{\theta_{\text{old}}}(s_t), -\epsilon, \epsilon),$$

and the clipped value loss can then be written as

$$L^{\text{V}}(\theta) = \frac{1}{2}\hat{\mathbb{E}}\Big[\max((V_\theta(s_t) - \hat{R}_t)^2, (V_\theta^{\text{clip}}(s_t) - \hat{R}_t)^2)\Big].$$

Furthermore, the third component of the loss is the *action entropy term*, given as

$$H(\theta) = \hat{\mathbb{E}}_t[H(\pi_\theta(\cdot|s_t)]. \tag{1}$$

The last term, the *proximal feature optimization term*, is given as

$$L^{\text{PFO}}(\theta) = \hat{\mathbb{E}}_t\Big[\|\phi_\theta(s_t) - \phi_{\theta_{\text{old}}}(s_t)\|_2^2\Big],$$

where $\|a\|_2^2$ denotes the sum of squares of the components of the vector a. Finally, the complete optimization objective function becomes

$$L(\theta) = L^{\text{CLIP}}(\theta) + c_{\text{V}}L^{\text{V}}(\theta) + c_{\text{E}}H(\theta) + c_{\text{PFO}}L^{\text{PFO}}(\theta). \tag{2}$$

In Table 6, we list all hyperparameters used in our implementation.

Table 6: Hyperparameters of our PPO implementation.

Hyperparameter	Role	Value
Transitions	Number of total training transitions	200,000,000[†]
Batch size	Batch size of experiences	4096
Steps	Number of transitions per environment in the data collection	128
Epochs	Number of epochs passing through collected data	3
Parallel environments	Number of environments in parallel during data collection	4096
Learning rate	Adam learning rate	$3 \cdot 10^{-4}$
β_1	Adam first moment factor	0.9
β_2	Adam second moment factor	0.999
ϵ_{Adam}	Constant added to the denominator of Adam update rule	10^{-8}
Max grad norm	Maximum norm of gradient	0.5
c_{V}	Value loss coefficient	0.5
c_{E}	Entropy coefficient	0.01
c_{PFO}	PFO coefficient	1.0
λ	GAE λ	0.95
ϵ	Clipping parameter	0.2

[†]In Space Invaders, we set the number of total transitions to 50,000,000, because of exceeding evaluation times.