# Reinforcement Learning for Finite Space Mean-Field Type Games

**Kai Shao, Jiacheng Shen, Mathieu Laurière**

**Keywords:** Deep Reinforcement Learning, Mean-Field Type Game

## Summary

Mean field type games (MFTGs) describe Nash equilibria between large coalitions: each coalition consists of a continuum of cooperative agents who maximize the average reward of their coalition while interacting non-cooperatively with a finite number of other coalitions. Although the theory has been extensively developed, we are still lacking efficient and scalable computational methods. Here, we develop reinforcement learning methods for such games in a finite space setting with general dynamics and reward functions. We start by proving that MFTG solution yields approximate Nash equilibria in finite-size coalition games. We then propose two algorithms. The first is based on quantization of mean-field spaces and Nash Q-learning. We provide convergence and stability analysis under suitable conditions. We then propose a deep reinforcement learning algorithm, which can scale to larger spaces. Numerical experiments in 4 environments with mean-field distributions of dimension up to 200 show the scalability and efficiency of the proposed method.

## Contribution(s)

1. We prove that the solution of an MFTG provides an $\epsilon$-Nash equilibrium for a game between finite-size coalitions (Theorem 2.4), which provides a motivation for solving MFTGs.
   **Context:** None
2. We propose a tabular RL method based on quantization of the continuous mean-field spaces and Nash Q-learning (Hu & Wellman, 2003). We prove the convergence of this algorithm under suitable conditions and analyze the error due to the discretization (Theorem 3.2).
   **Context:** None
3. We propose a deep RL algorithm based on DDPG (Lillicrap et al., 2016) which does not require quantization and hence is more scalable to problems with a large number of states.
   **Context:** None
4. We illustrate both methods in 4 environments with distribution in dimension up to 200.
   **Context:** Since this paper is the first to propose RL algorithms for (finite space) MFTGs with general dynamics and rewards, there is no standard baseline to compare with. We thus carry out a comparison with two baselines inspired by independent learning.

# Efficient Information Sharing for Training Decentralized Multi-Agent World Models

**Xiaoling Zeng, Qi Zhang**

**Keywords:** cooperative multi-agent reinforcement learning, decentralized world models, multi-agent communication.

## Summary

World models, which were originally developed for single-agent reinforcement learning, have recently been extended to multi-agent settings. Due to unique challenges in multi-agent reinforcement learning, agents' independent training of their world models often leads to underperforming policies, and therefore existing work has largely been limited to the centralized training framework that requires excessive communication. As communication is key, we ask the question of how the agents should communicate efficiently to train and learn policies from their decentralized world models. We address this question progressively. We first allow the agents to communicate with unlimited bandwidth to identify which algorithmic components would benefit the most from what types of communication. Then, we restrict the inter-agent communication with a predetermined bandwidth limit to challenge the agents to communicate efficiently. Our algorithmic innovations develop a scheme that prioritizes important information to share while respecting the bandwidth limit. The resulting method yields superior sample efficiency, sometimes even over centralized training baselines, in a range of cooperative multi-agent reinforcement learning benchmarks.

## Contribution(s)

1. This paper proposes a model-based MARL method that explicitly considers communication in both the world model and the actor-critic training stages, and analyzes the impact of communication bandwidth on decentralized training.
   **Context:** Previous work either builds on model-free MARL, discussing only experience sharing under different bandwidths with limited model and information diversity, or extracts shared agent information as features for centralized training, but omits these features during decentralized execution (Gerstgrasser et al., 2023; Venugopal et al., 2023).
2. Our work systematically studies information sharing under bandwidth limitations, aiming to optimize communication efficiency while guaranteeing performance within a decentralized framework.
   **Context:** Existing methods fail to effectively deal with communication bandwidth constraints and often rely on Euclidean distance constraints to filter communication neighbors (Toledo & Prorok, 2024).

# Recursive Reward Aggregation

**Yuting Tang** **Yivan Zhang** **Johannes Ackermann**
**Yu-Jie Zhang** **Soichiro Nishimori** **Masashi Sugiyama**

**Keywords:** Markov decision process, reward aggregation, policy preference, Bellman equation, algebraic data type, dynamic programming, recursion scheme, algebra fusion, bidirectional process

## Summary

In reinforcement learning (RL), agents typically learn desired behaviors by maximizing the (discounted) sum of rewards, making the design of reward functions crucial for aligning the agent behavior with specific objectives. However, because rewards often carry intrinsic meanings tied to the task, modifying them can be challenging and may introduce complex trade-offs in real-world scenarios. In this work, rather than modifying the reward function itself, we propose leveraging different reward aggregation functions to achieve different behaviors. By introducing an algebraic perspective on Markov decision processes (MDPs), we show that the Bellman equations naturally emerge from the recursive generation and aggregation of rewards. This perspective enables the generalization of the standard discounted sum to other recursive aggregation functions, such as discounted max and Sharpe ratio. We empirically evaluate our approach across diverse environments using value-based and actor-critic algorithms, demonstrating its effectiveness in optimizing a wide range of objectives. Furthermore, we apply our method to a real-world portfolio optimization task, showcasing its potential for practical deployment in decision-making applications where objectives cannot easily be expressed as the discounted sum of rewards.

## Contribution(s)

1. We provide an algebraic perspective on the recursive structure of MDPs based on fusion.
   **Context:** The algebra of recursive functions (Meijer et al., 1991; Bird & de Moor, 1997; Hutton, 1999) is a well-studied topic in functional programming. The fusion technique, explored by Hinze et al. (2010), has been applied to dynamic programming (Bellman, 1966; De Moor, 1994; Bertsekas, 2022). In the context of RL, the recursive structure of the discounted sum of rewards was studied by Hedges & Sakamoto (2022). Our diagrammatic representation of recursive reward generation and aggregation processes is inspired by Gavranović (2022).
2. We generalize the Bellman equations and Bellman operators for the standard discounted sum to other recursive aggregation functions, providing greater flexibility in goal specification.
   **Context:** The problem of alternative reward aggregations is not entirely new. Prior works have explored objectives such as optimizing the maximum (Quah & Quek, 2006; Gottipati et al., 2020; Veviurko et al., 2024), minimum (Cui & Yu, 2023), top-k (Wang et al., 2020), and Sharpe ratio (Nägele et al., 2024) of rewards. Specifically, the method proposed by Cui & Yu (2023) is a special case of our framework, where the recursive structure is on the original reward space, and the update function is order-preserving.
3. We extend existing RL algorithms by incorporating the generalized Bellman operators and empirically demonstrate their effectiveness across various tasks.
   **Context:** While our method modifies the Bellman operators within the base RL algorithms, the fundamental structures of Q-learning (Watkins, 1989; Watkins & Dayan, 1992), PPO (Schulman et al., 2017), and TD3 (Fujimoto et al., 2018) remain unchanged.

# A Finite-Sample Analysis of an Actor-Critic Algorithm for Mean-Variance Optimization in a Discounted MDP

**Tejaram Sangadi, Prashanth L.A., Krishna Jagannathan**

**Keywords:** Risk-Sensitive RL, Temporal Difference (TD) Learning, SPSA, Sample Complexity.

## Summary

In many practical applications of reinforcement learning (RL), such as finance and mobility, safety considerations are paramount. Rather than solely maximizing expected rewards, one must also account for risk to ensure reliable decision-making. Traditional RL primarily focuses on expected reward maximization, a well-studied paradigm with both empirical and theoretical breakthroughs. In this paper, we adopt an alternative approach that integrates risk-awareness into policy optimization. Despite extensive research in risk-neutral RL, analyzing risk-sensitive RL algorithms remains challenging, as each risk metric requires a distinct analytical framework. We focus on variance—an intuitive and widely used risk measure—and analyze the **M**ean-**V**ariance **S**imultaneous **P**erturbation **S**tochastic **A**pproximation **A**ctor-**C**ritic (MV-SPSA-AC) algorithm, establishing finite-sample theoretical guarantees for the discounted reward Markov Decision Process (MDP) setting. Our analysis covers both policy evaluation and policy improvement within the actor-critic framework. We study a Temporal Difference (TD) learning algorithm with linear function approximation (LFA) for policy evaluation and derive finite-sample bounds that hold in both the mean-squared sense and with high probability under tail iterate averaging, with and without regularization. Additionally, we analyze the actor update using a simultaneous perturbation-based approach and establish convergence guarantees. These results contribute to the theoretical understanding of risk-sensitive actor-critic methods in RL, offering insights into variance-based risk-aware policy optimization.

## Contribution(s)

1. We consider mean-variance optimization in a discounted MDP, and derive finite-sample guarantees for an actor-critic algorithm, with a critic based on linear function approximation, and an actor based on SPSA.
   **Context:** We consider a mean-variance MDP with the variance of the *return,* whose expectation is the usual risk-neutral objective. For this problem, existing work (L.A. & Ghavamzadeh, 2016) provides only asymptotic convergence guarantees.
2. For mean-variance policy evaluation, we employ TD learning with linear function approximation. We derive finite-sample bounds that hold (i) in the mean-squared sense and (ii) with high probability under tail iterate averaging, with and without regularization. Notably, our analysis for the regularized TD variant holds for a universal step size.
   **Context:** Non-asymptotic policy evaluation bounds are not available for variance of the return in a discounted MDP.
3. We employ an SPSA-based actor for policy optimization, and obtain an $O(n^{-1/4})$ bound in the number of actor iterations.
   **Context:** Notably, we resort to an SPSA-based actor, since the policy gradient theorem for variance is not amenable for direct use in an actor-critic algorithm; see L.A. & Ghavamzadeh (2016). Further, finite-sample bounds for a SPSA-based actor-critic algorithm are not available, even in the risk-neutral RL setting, to the best of our knowledge.

# Finite-Time Analysis of Minimax Q-Learning

**Narim Jeong, Donghwan Lee**

**Keywords:** Minimax Q-learning, finite-time analysis, control theory, switched systems.

## Summary

The goal of this paper is to present a finite-time analysis of minimax Q-learning and its smooth variant for two-player zero-sum Markov games, where the smooth variant is derived by using the Boltzmann operator. To the best of the authors' knowledge, this is the first work in the literature to provide such results. To facilitate the analysis, we introduce lower and upper comparison systems and employ switching system models. The proposed approach can not only offer a simpler and more intuitive framework for analyzing convergence but also provide deeper insights into the behavior of minimax Q-learning and its smooth variant. These novel perspectives have the potential to reveal new relationships and foster synergy between ideas in control theory and reinforcement learning.

## Contribution(s)

1. This paper presents a finite-time analysis of minimax Q-learning and its smooth variant with the Boltzmann operator, which is the first work to provide such results, as far as the authors are aware.
   **Context:** Most of the existing literature addresses its asymptotic convergence (Littman, 2001; Zhu & Zhao, 2020) or the convergence of the modified algorithms (Diddigi et al., 2022; Fan et al., 2019). Compared to others, our method can provide stronger convergence results through the finite-time analysis. Moreover, this paper addresses the vanilla minimax Q-learning and its smooth variant, which are based on the independently and identically distributed observations and the constant step-size in the tabular domain. Although we utilize these settings to simplify the analysis, our approach can be expanded to include more complex Markovian observation models by employing the methods described in Srikant & Ying (2019) and Bhandari et al. (2018).
2. By employing the switching system model for the convergence analysis, this paper contributes new insights into the convergence analysis of minimax Q-learning and the recently developed switching system framework for the finite-time analysis of Q-learning (Lee et al., 2022).
   **Context:** It is noteworthy to highlight that while the switching system model introduced in Lee et al. (2022) has been used as a basis, the main analysis and proof in this work significantly differed from those in Lee et al. (2022). In addition, we present a simulation result to empirically validate our method for the convergence analysis of the minimax Q-learning and its smooth variant that makes use of the switching system model.
3. This paper suggests the theoretically straightforward convergence analysis based on the control-theoretic concepts.
   **Context:** On the basis of the simple analytical approach, our analysis will help to reveal new relationships and promote mutual understanding between the control theory and RL.

# Im*pool*a: The Power of Average Pooling for Image-Based Deep Reinforcement Learning

**Raphael Trumpp, Ansgar Schäfftlein, Mirco Theile, Marco Caccamo**

**Keywords:** Network architecture, network scaling, image encoder, Procgen Benchmark

## Summary

As image-based deep reinforcement learning tackles more challenging tasks, increasing model size has become an important factor in improving performance. Recent studies achieved this by focusing on the parameter efficiency of scaled networks, typically using Impala-CNN, a 15-layer ResNet-inspired network, as the image encoder. However, while Impala-CNN evidently outperforms older CNN architectures, potential advancements in network design for deep reinforcement learning-specific image encoders remain largely unexplored. We find that replacing the flattening of output feature maps in Impala-CNN with global average pooling leads to a notable performance improvement. This approach outperforms larger and more complex models in the Procgen Benchmark, particularly in terms of generalization. We call our proposed encoder model Im*pool*a-CNN. A decrease in the network's translation sensitivity may be central to this improvement, as we observe the most significant gains in games without agent-centered observations. Our results demonstrate that network scaling is not just about increasing model size—efficient network design is also an essential factor.

## Contribution(s)

1. This work proposes the Impoola-CNN as image encoder for image-based deep reinforcement learning (DRL). Impoola-CNN is built upon the widely used Impala-CNN and enhances its network architecture by leveraging global average pooling (GAP).
   **Context:** The state-of-the-art Impala-CNN image encoder does not utilize GAP. In contrast, GAP is used in many popular network architectures in computer vision (He et al., 2016; Xie et al., 2017; Huang et al., 2017; Hu et al., 2018; Liu et al., 2022).
2. Our analysis for the full Procgen Benchmark demonstrates that Impoola-CNN excels at generalization, especially in environments without agent-centered observations. We show that Impoola-CNN outperforms other works on scaled networks in DRL.
   **Context:** The Procgen Benchmark (Cobbe et al., 2020) allows for the evaluation of generalization capabilities of image-based DRL agents, which is hard to assess in Atari games.
3. We identify reduced translation sensitivity in Impoola-CNN as a key distinction from Impala-CNN. Moreover, we find that Impoola-CNN is affected by fewer dormant neurons.
   **Context:** GAP reduces translation sensitivity (Lin, 2013) and is considered a strong inductive bias in computer vision. Sokar et al. (2023) identified the dormant neuron phenomenon, i.e., a large fraction of neurons yielding near-zero output during, as a cause of wide-ranging performance decrease in scaled networks in DRL.

# Fast Adaptation with Behavioral Foundation Models

**Harshit Sikchi, Andrea Tirinzoni, Ahmed Touati, Yingchen Xu, Anssi Kanervisto, Scott Niekum, Amy Zhang, Alessandro Lazaric, Matteo Pirotta**

**Keywords:** Unsupervised Learning, Self Supervised learning, Zero shot RL, Adaptation, Finetuning

## Summary

Unsupervised zero-shot reinforcement learning (RL) has emerged as a powerful paradigm for pretraining behavioral foundation models (BFMs), enabling agents to solve a wide range of downstream tasks specified via reward functions without additional test-time learning or planning. This is achieved by learning self-supervised task embeddings alongside corresponding near-optimal behaviors, and incorporating an inference procedure to directly retrieve the latent task embedding and associated policy for any given reward function. In this work, we demonstrate that existing unsupervised zero-shot RL pre-training methods discover a latent task embedding space containing more performant policies than those identified by their inference procedure, making them well-suited for fast adaptation. Motivated by this observation, we propose both actor-critic and actor-only fast adaptation strategies that search in the low-dimensional task-embedding space of the pre-trained BFM to rapidly improve the performance of its zero-shot policies on any downstream task. Notably, our approach mitigates the initial "unlearning" phase commonly observed when fine-tuning pre-trained RL models. We evaluate our fast adaptation strategies on top of four state-of-the-art zero-shot RL methods in multiple navigation and locomotion domains. Our results show that they achieve 10-40% improvement over their zero-shot performance in only a few episodes, outperforming existing baselines.

## Contribution(s)

1. We empirically investigate the task-representation space learned by a family of unsupervised zero-shot RL methods and show that it contains policies achieving significantly higher returns than the one output by the zero-shot inference procedure.
   **Context:** Prior works in zero-shot RL (Touati et al., 2023; Park et al., 2024; Agarwal et al., 2024) implicitly assume that zero-shot inference is the optimal way to prompt a pre-trained model for behaviors optimizing tasks specified by reward functions. We challenge such an assumption and show that this is not the case.
2. We propose two fast-adaptation algorithms: a) Residual Latent Adaptation (ReLA), an approach that optimizes for a policy in the BFM's task-representation space by training an additional smaller critic to estimate the cumulative reward not captured by the pre-trained BFM. b) Lookahead Latent Adaptation (LoLA), a computationally efficient approach that leverages policy gradients with lookahead returns without updating the pre-trained critic.
   **Context:** Prior approaches to adaptation either fine-tune the entire pre-trained critic and perform policy optimization in the action space (Nair et al., 2020; Nakamoto et al., 2023), or learn policy residuals (Silver et al., 2018; Johannink et al., 2019; Rana et al., 2023).
3. We evaluate our approaches on top of four state-of-the-art zero-shot RL methods in multiple navigation and locomotion domains, and show that they achieve 10-40% improvement over their zero-shot performance. Furthermore, we observe that our approach LoLA avoids the initial "unlearning" phase commonly observed in the literature.
   **Context:** Prior approaches for fine-tuning RL models without retaining training data (Luo et al., 2023; Zhou et al., 2024) observe a sharp decrease in performance due to distribution shift.

# Multi-Task Reinforcement Learning Enables Parameter Scaling

**Reginald McLean, Evangelos Chatzaroulas, J.K. Terry, Isaac Woungang, Nariman Farsad, Pablo Samuel Castro**

**Keywords:** Reinforcement learning, multi-task reinforcement learning, parameter scaling

## Summary

Multi-task reinforcement learning (MTRL) aims to endow a single agent with the ability to perform well on multiple tasks. Recent works have focused on developing novel sophisticated architectures to improve performance, often resulting in larger models; it is unclear, however, whether the performance gains are a consequence of the architecture design itself or the extra parameters. We argue that gains are mostly due to scale by demonstrating that naïvely scaling up a simple MTRL baseline to match parameter counts outperforms the more sophisticated architectures, and these gains benefit most from scaling the critic over the actor. Additionally, we explore the training stability advantages that come with task diversity, demonstrating that increasing the number of tasks can help mitigate plasticity loss. Our findings suggest that MTRL's simultaneous training across multiple tasks provides a natural framework for beneficial parameter scaling in reinforcement learning, challenging the need for complex architectural innovations.

## Contribution(s)

1. We perform a thorough empirical evaluation that shows that parameter scaling with simple architectures exceeds performance of complex MTRL-specific architectures on Meta-World benchmarks.
   **Context:** Increasing the scale of reinforcement learning models has been of particular interest of late (Obando Ceron et al., 2024; Nauman et al., 2024b; Schwarzer et al., 2023). This interest is likely due to the advances that other areas of research, such as computer vision (Zhai et al., 2022) and natural language processing (Kaplan et al., 2020), have made thanks to increasing the amount of data, compute, and model parameters during training. While some works have noted that naïvely scaling can be detrimental to performance (Obando Ceron et al., 2024), MTRL models have been quietly increasing in size. We compare results across recent multi-task reinforcement learning (MTRL) specific architectures, finding that increasing the parameter count on a simple feed forward baseline architecture outperforms the MTRL specific architectures.
2. We uncover a significant inverse relationship between task diversity and plasticity loss, showing that models trained on more tasks maintain neuron activation even as parameter counts increase.
   **Context:** When increasing the parameter scale alone, recent works have found that networks exhibit a loss of network plasticity (Lyle et al., 2023), or even a significant drop in performance when naïvely scaling (Obando Ceron et al., 2024). Across three different sizes of benchmark tasks, ten, twenty-five, and fifty tasks, we uncover a previously unknown relationship between the number of tasks and the number of parameters. We find that by increasing both model scale and the number of tasks being trained on can be an effective method to mitigate plasticity loss.
3. We empirically identify that critic scaling provides greater benefits than actor scaling in MTRL settings.
   **Context:** Recently, Nauman et al. (2024b) uncovered that the performance of the critic of an actor-critic algorithm was more affected by scale in single task settings. We extend this question to MTRL and have similar findings: that the scaling of the critic is more tightly coupled with the performance of the MTRL agent as a whole than the scale of the actor.

# Eau De $Q$-Network: Adaptive Distillation of Neural Networks in Deep Reinforcement Learning

**Théo Vincent** **Tim Faust** **Yogesh Tripathi**
**Jan Peters** **Carlo D'Eramo**

**Keywords:** Deep Reinforcement Learning, Sparse Training, Distillation.

## Summary

Recent works have successfully demonstrated that sparse deep reinforcement learning agents can be competitive against their dense counterparts. This opens up opportunities for reinforcement learning applications in fields where inference time and memory requirements are cost-sensitive or limited by hardware. To achieve a high sparsity level, the most effective methods use a dense-to-sparse mechanism where the agent's sparsity is gradually increased during training. Until now, those methods rely on hand-designed sparsity schedules that are not synchronized with the agent's learning pace. Crucially, the final sparsity level is chosen as a hyperparameter, which requires careful tuning as setting it too high might lead to poor performances. In this work, we address these shortcomings by crafting a dense-to-sparse algorithm that we name *Eau De Q-Network* (EauDeQN), where the online network is a pruned version of the target network, making the classical temporal-difference loss a distillation loss. To increase sparsity at the agent's learning pace, we consider multiple online networks with different sparsity levels, where each online network is trained from a shared target network. At each target update, the online network with the smallest loss is chosen as the next target network, while the other networks are replaced by a pruned version of the chosen network. Importantly, one online network is kept with the same sparsity level as the target network to slow down the distillation process if the other sparser online networks yield higher losses, thereby removing the need to set the final sparsity level. We evaluate the proposed approach on the Atari 2600 benchmark and the MuJoCo physics simulator. Without explicit guidance, EauDeQN reaches high sparsity levels while keeping performances high. We also demonstrate that EauDeQN adapts the sparsity schedule to the neural network architecture and the training length. Our code is publicly available at https://github.com/theovincent/EauDeDQN and the trained models are uploaded at https://huggingface.co/TheoVincent/Atari_EauDeQN.

## Contribution(s)

1. We introduce *Eau De Q-Network* (EauDeQN), a dense-to-sparse reinforcement learning framework capable of adapting the sparsity schedule at the agent's learning pace while maintaining high performance. As a result, EauDeQN *discovers* a final sparsity level. This means that EauDeQN avoids sparsity levels that are too high to yield high return and therefore removes the need to tune the final sparsity level.
   **Context:** Prior works in reinforcement learning consider hand-designed sparsity schedules and hard-coded final sparsity levels (Graesser et al., 2022). EauDeQN is composed of Distill $Q$-Network (also introduced in this work, resembling Ceron et al. (2024)), which is responsible for gradually pruning the network during training, and Adaptive $Q$-Network (Vincent et al., 2025b), which brings an adaptive behavior w.r.t. the agent's learning pace.

# Disentangling Recognition and Decision Regrets in Image-Based Reinforcement Learning

**Alihan Hüyük, A. Ryo Koblitz, Atefeh Mohajeri, Matthew Andrews**

**Keywords:** image-based reinforcement learning, observational overfitting, over-specific representations, under-specific representations, recognition regret, decision regret

## Summary

In image-based reinforcement learning (RL), policies usually operate in two steps: first extracting lower-dimensional features from raw images (the "recognition" step), and then taking actions based on the extracted features (the "decision" step). Extracting features that are spuriously correlated with performance or irrelevant for decision-making can lead to poor generalization performance, known as *observational overfitting* in image-based RL. In such cases, it can be hard to quantify how much of the error can be attributed to poor feature extraction vs. poor decision-making. To disentangle the two sources of error, we introduce the notions of *recognition regret* and *decision regret*. Using these notions, we characterize and disambiguate the two distinct causes behind observational overfitting: *over-specific representations*, which include features that are not needed for optimal decision-making (leading to high decision regret), vs. *under-specific representations*, which only include a limited set of features that were spuriously correlated with performance during training (leading to high recognition regret). Finally, we provide illustrative examples of observational overfitting due to both over-specific and under-specific representations in maze environments and the Atari game Pong.

## Contribution(s)

1. We define recognition and decision regrets, which disentangle the regret induced by poor recognition policies vs. poor decision policies.
   **Context:** In image-based RL, most agents first extract features from images and then take actions based on the extracted features. When an RL agent does not perform well (measured by its regret), it is hard to tell whether this is due to a failure to extract features or a failure to plan good actions. Our definitions break down overall regret into two components that attribute it to one of these two failure modes.
2. By analyzing generalization performance through recognition and decision regrets, we characterize over-specific and under-specific representations as two distinct modes of observational overfitting in image-based RL.
   **Context:** Observational overfitting is a phenomenon in image-based RL that an agent learns to rely on information in the image that does not actually constitute a part of the environment state (like decorative elements in a video game) (Song et al., 2020). Using the notions of recognition and decision regrets, we identify this phenomenon can occur in two distinct forms: learning over-specific representations that include irrelevant features (not informative of the actual state) vs. under-specific representations that only include spurious features (correlated with the actual state but only during training, the correlations do not generalize).
3. We provide illustrative examples of observational overfitting due to both over-specific and under-specific representations in maze environments as well as the Atari game Pong.
   **Context:** These two environments contrast each other well in terms of their dimensionality, and they are commonly considered when exploring generalization in RL (e.g. Sonar et al., 2021; Taiga et al., 2023).

# Learning to Explore in Diverse Reward Settings via Temporal-Difference-Error Maximization

**Sebastian Griesbach, Carlo D'Eramo**

**Keywords:** Deep RL, Exploration, TD-error Maximization

## Summary

Numerous heuristics and advanced approaches have been proposed for exploration in different settings for deep reinforcement learning. Noise-based exploration generally fares well with dense-shaped rewards and bonus-based exploration with sparse rewards. However, these methods usually require additional tuning to deal with undesirable reward settings by adjusting hyperparameters and noise distributions. Rewards that actively discourage exploration can pose a major challenge. This is the case if the reward function contains an action cost and no other dense signal to follow. We propose a novel exploration method, Stable Error-seeking Exploration (SEE), that is robust across dense, sparse, and exploration-adverse reward settings. To this endeavor, we revisit the idea of maximizing the TD-error as a separate objective. Our method introduces three design choices to mitigate instability caused by: (i.) far off-policy learning, when a behavior policy is too out of distribution w.r.t. the target policy; (ii.) the conflict of interest of terminating an episode while the exploration objective follows an always positive reward signal; (iii.) the non-stationary nature of the TD-error as a target. SEE can be combined with off-policy algorithms without modifying the optimization pipeline of the original objective. In our experimental analysis, we show that a Soft-Actor Critic agent with the addition of SEE performs robustly across three diverse reward settings in a variety of tasks without hyperparameter adjustments.

## Contribution(s)

1. We propose methodological approaches to tackle three identified causes of instability as mentioned above: (1.) combining the exploration and exploitation policies into a single behavior policy to bridge the far-off-policy gap; (2.) using a maximum reward update, which is agnostic towards the length of an episode; (3.) conditioning the exploration value function on current estimates of the exploitation value function to inform it about the cause of change in the TD-error target.
   **Context:** Prior works investigated framing exploration as a separate optimization problem (Whitney et al., 2021; Schäfer et al., 2022) and maximizing the TD-error (Simmons-Edler et al., 2020). The combination of both is notoriously challenging to stabilize without altering the original optimization objective.
2. We incorporate the proposed solutions resulting in SEE, and combine it with well-established off-policy reinforcement algorithms, namely SAC (Tuomas Haarnoja et al., 2018) and TD3 (Scott Fujimoto et al., 2018). Our results show that the addition of SEE maintains performance in dense reward settings and improves robustness in sparse and exploration-adverse settings, without additional hyperparameter tuning.
   **Context:** We empirically compare the performances of the base algorithms and their SEE extensions across a set of environments with variants for dense, sparse, and exploration-adverse rewards.

# Nonparametric Policy Improvement in Continuous Action Spaces via Expert Demonstrations

**Agustin Castellano, Sohrab Rezaei, Jared Markowitz, Enrique Mallada**

**Keywords:** Policy Optimization, Policy Improvement, Imitation Learning, Nonparametric methods.

## Summary

The policy improvement theorem is a fundamental building block of classical reinforcement learning for discrete action spaces. Unfortunately, the lack of an analogous result for continuous action spaces with function approximation has historically limited the ability of policy optimization algorithms to make large step updates, undermining their convergence speed. Here we introduce a novel nonparametric policy that relies purely on data to take actions and that admits a policy improvement theorem for deterministic Markov Decision Processes (MDPs). By imposing mild regularity assumptions on the optimal policy, we show that, when data come from expert demonstrations, one can construct a nonparametric lower bound on the value of the policy, thus enabling its robust evaluation. The constructed lower bound naturally leads to a simple improvement mechanism, based on adding more demonstrations. We also provide conditions to identify regions of the state space where additional demonstrations are needed to meet specific performance goals. Finally, we propose a policy optimization algorithm that ensures a monotonic improvement of the lower bound and leads to high probability performance guarantees. These contributions provide a foundational step toward establishing a rigorous framework for policy improvement in continuous action spaces.

## Contribution(s)

*i*) We present a novel framework for nonparametric policies on continuous state and action spaces that only requires data coming from expert trajectories.
**Context:** Modern RL algorithms usually learn a parametrized policy (Schulman et al., 2017), a model of the environment, or both (Hafner et al., 2019; Janner et al., 2019).

*ii*) Robust policy evaluation: Under mild assumptions on the MDP, we can readily construct a lower bound on the optimal $Q$-function. Our policy is *greedy* with respect to this bound and surprisingly improves upon it.
**Context:** The expression for this lower bound ensures that greedy actions can be carried out in closed form, making our policy easy to implement and evaluate. In contrast, standard policy iteration (Sutton & Barto, 2018) relies on computing an (approximate) value function estimate of a policy.

*iii*) Policy improvement: Our framework leads to a policy improvement mechanism, in which more data yields ever tighter lower bounds. As a result, our policy sequentially improves on the new data.
**Context:** We provide sufficient conditions for our policy to be *strictly* improving on the new data points. Notably, this method allows for large policy updates, in contrast to policy gradient (Sutton et al., 1999) or trust region methods (Schulman et al., 2015), which take small enough steps to ensure improvement on average.

*iv*) Policy optimization with guarantees: We present a novel algorithm, inspired by minorization maximization, that monotonically improves our lower value estimate, leading to high probability performance guarantees.
**Context:** We derive easy-to-check conditions (based on the value function bounds and sampled states) that either guarantee a certain suboptimality or suggest a location where new demonstrations are necessary to meet the performance requirements.

# DisDP: Robust Imitation Learning via Disentangled Diffusion Policies

**Pankhuri Vanjani, Paul Mattes, Xiaogang Jia, Vedant Dave, Rudolf Lioutikov**

**Keywords:** Imitation learning, Diffusion policy, Multi-View Disentanglement

## Summary

This work introduces Disentangled Diffusion Policy (DisDP), an Imitation Learning (IL) method that enhances robustness by integrating multi-view disentanglement into diffusion-based policies. For robots to be deployed on a large scale across various applications they have to be robust against different perturbations, including sensor noise, complete sensor dropout and environmental variations. Existing IL methods struggle to generalize under such conditions, as they typically assume consistent, noise-free inputs. To address this limitation, DisDP structures sensory inputs into shared and private representations, preserving task-relevant global features while retaining distinct details from individual sensors. Additionally, Disentangled Behavior Cloning (DisBC) is introduced, a disentangled Behavior Cloning (BC) policy, to demonstrate the general applicance of disentanglement for IL. This structured representation improves resilience against sensor dropouts and perturbations. Evaluations on The Colosseum and Libero benchmarks demonstrate that disentangled policies achieve better performance in general and exhibit greater robustness to any perturbations compared to their baseline policies.

## Contribution(s)

1. *Introducing Disentangled Diffusion Policy (DisDP), an Imitation Learning (IL) method that improves robustness to sensor noise and dropouts by structuring sensor inputs into shared and private representations.*

   **Context:** Prior IL methods rely on consistent, noise-free sensor inputs, which limits their effectiveness in real-world scenarios.

2. *Leveraging Multi-View Disentanglement for enhancing robustness and interpretability of the behavior policies.*

   **Context:** This implementation displays the general application of Multi-view disentanglement for robot policies. It uses contrastive and orthogonality constraints to separate shared and unique information. This approach enhances the interpretability by visualizing the shared and private representations.

3. *Provides an extensive experimental analysis on the effect of the Colosseum and Libero benchmarks in sensor failure and environment perturbation scenarios.*

   **Context:** This work shows the performance degradation of behavior policies with unreliable sensors and environmental variations. Additionally, it shows how disentangled latent spaces benefit in these scenarios.

# Mitigating Goal Misgeneralization via Minimax Regret

**Karim Abdel Sadek[=], Matthew Farrugia-Roberts[=], Usman Anwar, Hannah Erlebach, Christian Schroeder de Witt, David Krueger, Michael Dennis**

**Keywords:** Goal Misgeneralization, Unsupervised Environment Design, AI Safety, AI Alignment.

## Summary

Safe generalization in reinforcement learning requires not only that a learned policy *acts capably* in new situations, but also that it uses its capabilities *towards the pursuit of the designer's intended goal.* The latter requirement may fail when a *proxy goal* incentivizes similar behavior to the intended goal within the training environment, but not in novel deployment environments. In this setting, policies may behave as if in pursuit of the proxy goal in deployment—a phenomenon known as *goal misgeneralization.* In this paper, we theoretically investigate the possibility of goal misgeneralization under *maximum expected value (MEV)* and *minimax expected regret (MMER)* objectives, and empirically validate our results. Our findings underscore minimax expected regret as a promising principle for mitigating goal misgeneralization.

## Contribution(s)

1. We introduce a problem setting called a *proxy-distinguishing distribution shift,* capturing a class of situations in which goal misgeneralization can be elicited and studied.
   **Context:** In a proxy-distinguishing distribution shift, optimizing a given proxy goal also optimizes the true goal in most training situations, but optimizing the proxy goal can be suboptimal under the true goal in most deployment situations (in particular, so-called *distinguishing levels*). We do not assume training methods have knowledge of the proxy goal.
2. We prove that, under a proxy-distinguishing distribution shift, approximately maximizing expected value on the training distribution permits a misgeneralizing solution if the proportion of distinguishing levels in the training distribution is low enough (Theorem 1).
   **Context:** *Exactly* maximizing expected value on the training distribution permits misgeneralization if *no* distinguishing levels are seen in training. We model *possible* goal misgeneralization; *actual* goal misgeneralization also depends on the agent's inductive biases.
3. We prove that, under a proxy-distinguishing distribution shift, no approximate solution of the minimax expected regret objective exhibits goal misgeneralization (Theorem 2).
   **Context:** Theorem 2 holds for fully observable environments; we include a generalization to partially observable environments in the supplementary materials (Theorem 3).
4. Experiments suggest (no statistical significance analysis) that, under conditions approximating a proxy-distinguishing distribution shift in procedurally generated grid-world environments, policies learned using MEV-based training exhibit goal misgeneralization when the proportion of distinguishing levels in the training distribution is low enough (§7.1).
   **Context:** Langosco et al. (2022) demonstrated goal misgeneralization with zero distinguishing levels, we extend this finding to the case with a small positive proportion.
5. Experiments suggest (no statistical significance analysis) that, under the same conditions, existing regret-based unsupervised environment design (UED) methods, PLR$^{\perp}$ (Jiang et al., 2021a) and ACCEL (Parker-Holder et al., 2022), (1) can detect rare distinguishing levels and increase their proportion in the training distribution, and (2) are more robust to goal misgeneralization than MEV-based training is (§7.2).
   **Context:** In some cases, less advanced UED methods fail to find MMER policies, and still exhibit goal misgeneralization (§7.3, §7.4), indicating that more mature UED methods are needed to achieve the potential of MMER for preventing goal misgeneralization in practice.

# Long-Horizon Planning with Predictable Skills

**Nico Gürtler, Georg Martius**

**Keywords:** model-based reinforcement learning, skill learning, long-horizon planning, long-term credit assignment, compounding model errors

## Summary

Model-based reinforcement learning (RL) leverages learned world models to plan ahead or train in imagination. Recently, this approach has significantly improved sample efficiency and performance across various challenging domains ranging from playing games to controlling robots. However, there are fundamental limits to how accurate the long-term predictions of a world model can be, for example due to unstable environment dynamics or partial observability. These issues are further exacerbated by the compounding error problem. Model-based RL is therefore generally limited to short rollouts with the world model, and consequently struggles with long-term credit assignment. We argue that this limitation can be addressed by modeling the outcome of temporally extended skills instead of the effect of primitive actions. To this end, we propose a mutual-information-based skill learning objective that ensures predictable, diverse, and task-related behavior. The resulting skills compensate for perturbations and drifts, enabling stable long-horizon planning. We thus introduce *Stable Planning with Temporally Extended Skills (SPlaTES)*, a sample-efficient hierarchical agent consisting of model predictive control with an abstract skill world model on the higher level, and skill execution on the lower level.

## Contribution(s)

1. We introduce SPlaTES, a sample-efficient hierarchical RL algorithm that learns temporally extended skills on the lower level, and an abstract world model predicting skill outcomes on the higher level. Both levels are model-based and perform model predictive control over different timescales.
   **Context:** Existing model-based hierarchical agents either do not use an abstract world model for planning (Hafner et al., 2022), are restricted to a pre-defined symbolic abstraction of the environment (Achterhold et al., 2023), or require a pre-collected dataset with high-quality skill behavior (Shi et al., 2023) for offline learning.
2. We show that our mutual-information-based skill learning objective results in diverse and predictable skill outcomes. The temporal extent of the skills enables error-correcting behavior contributing to the stability of the high-level dynamics.
   **Context:** Like Gregor et al. (2016) and Achterhold et al. (2023), we consider the mutual information of the skill outcome and skill vector. However, we show empirically that transforming such a sparse reward into a dense one is crucial for obtaining good performance. We furthermore condition on the intra-skill time step and start state to enable robust error compensation.
3. Planning over entire episodes enables SPlaTES to solve challenging long-horizon tasks without resorting to temporal difference learning for long-term credit assignment. Our empirical evaluation shows that SPlaTES outperforms competitive skill-based and model-based baselines.
   **Context:** Distilling the behavior of the hierarchical agent into a flat TD-MPC2 model (Hansen et al., 2024) results in decreased performance and myopic behavior. We conclude that SPlaTES performs long-term credit assignment on time scales that are difficult to achieve with non-hierarchical temporal difference learning.

# HANQ: Hypergradients, Asymmetry, and Normalization for Fast and Stable Deep $Q$-Learning

**Braham Snyder, Chen-Yu Wei**

**Keywords:** off-policy reinforcement learning (RL), offline RL, temporal difference learning, bootstrapping, instability, return degradation, value estimation

## Summary

In reinforcement learning (RL), deep $Q$-learning algorithms are often more sample- and compute-efficient than alternatives like the Monte Carlo policy gradient, but tend to suffer from instability that limits their use in practice. Some of this instability can be mitigated through a *target network*, yet this doubles memory usage and arguably slows down convergence. In this work, we explore the possibility of stabilization (returns do not drop with further gradient steps) without sacrificing the speed of convergence (high returns do not require many gradient steps). Inspired by self-supervised learning (SSL) and adaptive optimization, we empirically arrive at three modifications to the standard deep $Q$-network (DQN) — no two of which work well alone in our ablations. These modifications are, in the order of our experiments: 1) an **A**symmetric *predictor* in the neural network, 2) a particular combination of **N**ormalization layers, and 3) **H**ypergradient descent on the learning rate. Aligning with prior work in SSL, **HANQ** (pronounced "*hank*") avoids DQN's target network, uses the same number of hyperparameters as DQN, and yet matches or exceeds DQN's performance in our offline RL experiments on three out of four environments.

## Contribution(s)

1. We propose to replace the target network in deep $Q$-network (DQN) with an asymmetric predictor and normalization layers to stabilize training. Empirical results suggest the promise of our approach given appropriate learning rate tuning.
   **Context:** Asymmetric architectures have been explored in self-supervised learning (Grill et al., 2020; Chen & He, 2021) and reinforcement learning (RL) (Gelada et al., 2019; Pitis et al., 2020; Guo et al., 2022; Liu et al., 2022; Tang et al., 2023; Wang, 2024; Eysenbach et al., 2024; Amortila et al., 2024; Myers et al., 2025). However, to our knowledge, all prior RL works study auxiliary losses or goal-based RL, and typically keep the target network and increase the number of hyperparameters. We study pure end-to-end reward maximization, we remove the target network, and we do not increase the number of hyperparameters.
2. Noting that promise of our first contribution, we use hypergradient descent for that tuning, which achieves more stability without compromising the convergence rate in our experiments: our algorithm (HANQ) matches or outscores DQN in three of four environments.
   **Context:** We run offline RL experiments on three classic control environments and one Atari environment. Further, prior works investigate hypergradients for temporal difference learning (Sutton, 2022; Farahmand & Ghavamzadeh, 2021; Bedaywi et al., 2024; Javed et al., 2024). However, we find using hypergradient descent alone (or asymmetry alone) scores poorly.
3. Our extensive ablations suggest each component of HANQ is important for its high scores.
   **Context:** Prior works (Gallici et al., 2024; Elsayed et al., 2024) show normalization layers can often replace the stabilization benefit of a target network, but HANQ scores up to twice as high as PQN (Gallici et al., 2024).

# Benchmarking Massively Parallelized Multi-Task Reinforcement Learning for Robotics Tasks

**Viraj Joshi, Zifan Xu, Bo Liu, Peter Stone, Amy Zhang**

**Keywords:** Multi-Task Learning, Reinforcement Learning, Robotics.

## Summary

Multi-task Reinforcement Learning (MTRL) has emerged as a critical training paradigm for applying reinforcement learning (RL) to a set of complex real-world robotic tasks, which demands a generalizable and robust policy. At the same time, *massively parallelized training* has gained popularity, not only for significantly accelerating data collection through GPU-accelerated simulation but also for enabling diverse data collection across multiple tasks by simulating heterogeneous scenes in parallel. However, existing MTRL research has largely been limited to off-policy methods like SAC in the low-parallelization regime. MTRL could capitalize on the higher asymptotic performance of on-policy algorithms, whose batches require data from current policy, and as a result, take advantage of massive parallelization offered by GPU-accelerated simulation. To bridge this gap, we introduce a massively parallelized **M**ulti-**T**ask **Bench**mark for robotics (MTBench), an open-sourced benchmark featuring a broad distribution of 50 manipulation tasks and 20 locomotion tasks, implemented using the GPU-accelerated simulator IsaacGym. MTBench also includes four base RL algorithms combined with seven state-of-the-art MTRL algorithms and architectures, providing a unified framework for evaluating their performance. Our extensive experiments highlight the superior speed of evaluating MTRL approaches using MTBench, while also uncovering unique challenges that arise from combining massive parallelism with MTRL.

## Contribution(s)

1. This paper introduces MTBench, a unified GPU-accelerated benchmark for massively parallelized multi-task reinforcement learning (MTRL) in two robotics settings, manipulation and locomotion.
   **Context:** Existing robotics MTRL benchmarks, such as Meta-World (Yu et al., 2021), have impractically long experimental runtimes, hindering the development and reproducibility of MTRL research. Other GPU-accelerated benchmarks for robotics do not support MTRL out of the box. We address both of these concerns with our end-to-end MTRL benchmark.
2. This paper conducts comprehensive experiments to evaluate all aspects of MTRL, including base RL algorithms, gradient manipulation methods, and neural network architectures.
   **Context:** We confirm whether the reliance on off-policy methods in the MTRL literature holds in the massively parallel regime, and then evaluate a suite of MTRL schemes using on-policy methods across our evaluation settings.
3. This paper presents four key observations on applying existing MTRL schemes to massively parallelized training in robotics. These insights guide the selection of MTRL schemes and inform future research directions.
   **Context:** Massively parallelized training is emerging as a popular paradigm, introducing unique challenges for existing RL methods (D'Oro et al., 2022; Li et al., 2023; Gallici et al., 2024; Singla et al., 2024). However, MTRL development has yet to leverage this paradigm.

# Optimal discounting for offline Input-Driven MDP

**Randy Lefebvre, Audrey Durand**

**Keywords:** Offline RL, Input-Driven MDP, Bias-variance tradeoff, Discount factor

## Summary

Offline reinforcement learning has gained a lot of popularity for its potential to solve industry challenges. However, real-world environments are often highly stochastic and partially observable, leading long-term planners to overfit to offline data in model-based settings. Input-Driven Markov Decision Processes (IDMDPs) offer a way to work with some of the uncertainty by letting designers separate what the agent has control over (states) from what it cannot (inputs) in the environnement. These stochastic external inputs are often difficult to model. Under the assumption that the input model will be imperfect, we investigate the bias-variance tradeoff under shallow planning in IDMDPs. Paving the way to input-driven planning horizons, we also investigate the similarity of optimal planning horizons at different inputs given the structure of the input space.

## Contribution(s)

1. We provide new insights connecting the input structure to the state-value function in Input-Driven MDPs (Lemma 1).
   **Context:** This result is also applicable to MDPs and therefore generalizes the value function variation from Jiang et al. (2016) to any policy and any pair of states.
2. We provide a novel bound on the variance due to the error in the input model and the planning horizon in offline Input-Driven MDPs (Lemma 2), which we use to obtain the first existing bound on the planning loss for Exo-MDPs (Theorem 1).
   **Context:** Prior results (Jiang et al., 2015; Lefebvre & Durand, 2025) study the variance due to the error in the state model in a MDP, i.e. considering variables that the agent can control (whereas the agent cannot control the inputs).
3. We provide the first results on the optimal input-dependent discount factor in Input-Driven MDPs. We connect the planning loss at different inputs to the input structure (Lemma 3), allowing to control the variation of optimal input-dependent discount factors over the input space using the input structure (Theorem 2).
   **Context:** This connects to the (limited) work on state-dependent discount factors, focusing on the impact of the non-controllable variables (inputs) on the optimal planning horizon.

# Make the Pertinent Salient: Task-Relevant Reconstruction for Visual Control with Distractions

**Kyungmin Kim, JB Lanier, Roy Fox**

**Keywords:** Visual Control, Robust Representation Learning, Model-Based RL.

## Summary

Model-Based Reinforcement Learning (MBRL) has shown promise in visual control tasks due to its data efficiency. However, training MBRL agents to develop generalizable perception remains challenging, especially amid visual distractions that introduce noise in representation learning. We introduce Segmentation Dreamer (SD), a framework that facilitates representation learning in MBRL by incorporating a novel auxiliary task. Assuming that task-relevant components in images can be easily identified with prior knowledge in a given task, SD uses segmentation masks on image observations to reconstruct only task-relevant regions, reducing representation complexity. SD can leverage either ground-truth masks available in simulation or potentially imperfect segmentation foundation models. The latter is further improved by selectively applying the image reconstruction loss to mitigate misleading learning signals from mask prediction errors. In modified DeepMind Control suite and Meta-World tasks with added visual distractions, SD achieves significantly better sample efficiency and greater final performance than prior work and is especially effective in sparse reward tasks that had been unsolvable by prior work. In a real-world robotic lane-following task, our method trained with intentional distractions provides early evidence that a model-based method can transfer from simulation to a real robot under visual domain randomization.[a]

[a] Project page: https://indylab.github.io/SD

## Contribution(s)

1. This paper introduces a novel auxiliary task in model-based reinforcement learning (MBRL) to enhance representation learning in visually distracting environments. Our approach reconstructs control-relevant components while filtering out distractions, ensuring that latent embeddings focus on essential features.
   **Context:** While our method requires prior knowledge of task-relevant components, identifying these components is typically straightforward for practitioners in many robotics applications. Prior work using reconstruction-free auxiliary tasks relies on large amounts of data to infer important features, making them less sample-efficient.
2. This paper integrates segmentation foundation models to guide feature learning in visual control through task-relevant reconstruction targets, without incurring extra test-time overhead and while improving robustness to segmentation errors. This demonstrates an effective way to harness advances in computer vision for visual control tasks.
   **Context:** Prior approaches typically use segmentation models for input preprocessing, which adds deployment overhead and increases sensitivity to segmentation errors.
3. Our method learns effective visual control policies in environments with distractions, demonstrating success in DMC, where locomotion control requires handling contact dynamics; Meta-World, which involves robotic manipulation, occlusions, and multi-object interactions; and DuckieTown, where transferring lane-following behavior from simulation to reality must account for diverse perturbations, including foreground distractions.
   **Context:** Our method is sample-efficient, achieves record final performance, and is the only method capable of learning with sparse rewards in DMC.

# Reinforcement Learning for Human-AI Collaboration via Probabilistic Intent Inference

**Yuxin Lin, Seyede Fatemeh Ghoreishi, Tian Lan, Mahdi Imani**

**Keywords:** Reinforcement Learning, Probabilistic Intent Inference, Human-AI Collaboration, Belief-Space Planning, Decision-Making Under Uncertainty.

## Summary

Effective collaboration between humans and AI agents is increasingly essential as autonomous systems take on critical roles in domains like disaster response, healthcare, and robotics. However, achieving robust human-AI collaboration remains challenging due to the uncertainty, complexity, and unpredictability of human behavior, which is often difficult to convey explicitly to AI agents. This paper presents a belief-space reinforcement learning framework that enables AI agents to implicitly and probabilistically infer latent human intentions from behavioral data and integrate this understanding into robust decision-making. Our approach models human behavior at both the action (low) and subtask (high) levels, combining these with human and agent state information to construct a comprehensive belief state for the AI agent. We demonstrate that this belief state follows the Markov property, enabling the derivation of an optimal Bayesian policy under human and task uncertainty. Deep reinforcement learning is used to train an offline Bayesian policy across a wide range of human and task uncertainties, allowing real-time deployment to support effective human-AI collaboration. Numerical experiments demonstrate the effectiveness of the proposed policy in terms of cooperation, adaptability, and robustness.

## Contribution(s)

1. We develop a decision-making framework that represents the human behavioral model at two levels—low-level actions and high-level subtasks—allowing the AI agent to anticipate long-term human goals and adapt to changing task priorities in real-time.
   **Context:** Unlike prior models that focus on human rationality at a single (action) level, our approach incorporates hierarchical intent modeling, enhancing goal-aware human-AI collaboration and improving adaptability to dynamic environments.
2. We propose a structured belief state that captures state information alongside the posterior distribution of human intent, serving as a sufficient statistic for optimal Bayesian decision-making in human-AI collaboration.
   **Context:** Unlike existing Partially Observable Markov Decision Process (POMDP)-based frameworks that maintain beliefs over partially observable states, our belief state explicitly models uncertainty in high-level human intent, leading to more informed and adaptive decision-making under uncertainty.
3. We develop a deep reinforcement learning (DRL) approach that optimizes the AI agent's decision-making over the belief space, enabling dynamic adaptation to inferred human intent for effective long-term human-AI collaboration.
   **Context:** Unlike existing methods that optimize AI agents for pre-specified human tasks or rely on explicit feedback, our approach leverages a belief-space policy trained on human behaviors. This policy captures the AI's belief about human intent—including uncertainty in their goals and actions (theory of mind)—to optimize decision-making accordingly. This enables efficient real-time adaptation without requiring explicit human feedback.

# PufferLib 2.0: Reinforcement Learning at 1M steps/s

**Joseph Suarez**

**Keywords:** PufferLib, Reinforcement Learning, Library, Tools

## Summary

PufferLib is an open-source reinforcement learning project built around efficient and broadly compatible simulation. Our first-party suite of 12 environments each run at 1M steps/second. For existing environments, PufferLib provides one-line wrappers that eliminate common compatibility problems and fast vectorization to accelerate training. With PufferLib, you can use familiar libraries like CleanRL and SB3 to scale from classic benchmarks like Atari and Procgen to complex simulators like NetHack and Neural MMO 3. Code, documentation, demos, and less formal blog coverage are available at puffer.ai.

## Contribution(s)

1. One-line wrappers that make complex environments like Nethack, Neural MMO, Griddly, etc. compatible with any RL library that supports standard Gymnasium/PettingZoo formats.
   **Context:** Gymnasium and PettingZoo are the most widely used environment formats. This means PufferLib is compatible with the vast majority of environments using only a 1-line wrapper.
2. Drop-in vectorization for simulating environments in parallel. Most environments will see at least a 30% speed boost and 50%-3x with pooling. This is a broadly compatible contribution applicable to nearly all environments.
   **Context:** Gymnasium provides the most common vectorization backend. It is slow for the reasons outlined in the paper.
3. Puffer Ocean, a suite of 12 environments written in C that each simulate at >1M steps/second on a single CPU core.
   **Context:** A few of these have built-in AI opponents that can slow performance when search depth is increased. Base speed is >1M steps/second on a high end desktop core. Some environments run 10M steps/second.
4. A PPO demo that trains Ocean environments at 300k-1.2M steps/second on a single RTX 4090. Our standard architectures are MLP-LSTM or CNN-LSTM from 150k-1M parameters.
   **Context:** It's compatible with all third-party environments. Training is up to 30k steps/second on Atari, which is still 30x faster than the original CleanRL.

# Uncovering RL Integration in SSL Loss: Objective-Specific Implications for Data-Efficient RL

**Ömer Veysel Çağatan, Barış Akgün**

**Keywords:** Data Efficient RL, Self Predictive RL, Self Supervised Learning

## Summary

This paper presents a systematic analysis of the role of self-supervised learning (SSL) objectives and their modifications in data-efficient reinforcement learning. We investigate previously undocumented modifications in the Self-Predictive Representations (SPR) (Schwarzer et al., 2020) framework that significantly impact agent performance. We demonstrate that feature decorrelation-based SSL objectives can achieve comparable performance without relying on domain-specific modifications and show that the impact of these modifications persists even in more advanced models.

By conducting extensive experiments on the Atari 100k benchmark and DeepMind Control Suite, we provide insights into how different SSL objectives and their modifications affect learning efficiency across diverse environments. Our findings reveal that the choice and adaptation of SSL objectives play a crucial role in achieving data efficiency in self-predictive reinforcement learning, with implications for the design of future algorithms in this space.

## Contribution(s)

1. We demonstrate that previously undocumented SSL modifications in SPR (Schwarzer et al., 2020) - terminal state masking and prioritized replay weighting - are crucial for performance, with their removal leading to an 18% decrease in IQM score on Atari 100k
   **Context:** These modifications were silently adopted by subsequent work (D'Oro et al., 2023; Nikishin et al., 2022; Schwarzer et al., 2023) and their impact was not previously analyzed
2. We show that the Barlow Twins SSL objective (Zbontar et al., 2021) can come within 5% of SPR's performance without using domain-specific modifications, and VICReg (Bardes et al., 2021) can match PlayVirtual's (Yu et al., 2021) performance in continuous control tasks.
   **Context:** Prior work on SSL in reinforcement learning relied heavily on problem-specific modifications to achieve strong performance (Schwarzer et al., 2020; D'Oro et al., 2023; Schwarzer et al., 2023).
3. We establish that the impact of SSL modifications remains proportionally consistent in more sophisticated models, with unmodified versions of SR-SPR and BBF showing similar relative performance degradation despite having base IQM scores 3x and 2x higher than SPR, respectively.
   **Context:** Previous work on SR-SPR (D'Oro et al., 2023; Nikishin et al., 2022) and BBF (Schwarzer et al., 2023) did not investigate the role of these modifications in their improved performance.

# Benchmarking Partial Observability in Reinforcement Learning with a Suite of Memory-Improvable Domains

**Ruo Yu Tao , Kaicheng Guo , Cameron Allen , George Konidaris**

**Keywords:** reinforcement learning, partial observability, benchmarking

## Summary

Mitigating partial observability is a necessary but challenging task for general reinforcement learning algorithms. To improve an algorithm's ability to mitigate partial observability, researchers need comprehensive benchmarks to gauge progress. Most algorithms tackling partial observability are only evaluated on benchmarks with simple forms of state aliasing, such as feature masking and Gaussian noise. Such benchmarks do not represent the many forms of partial observability seen in real domains, like visual occlusion or unknown opponent intent. We argue that a partially observable benchmark should have two key properties. The first is coverage in its forms of partial observability, to ensure an algorithm's generalizability. The second is a large gap between the performance of a agents with more or less state information, all other factors roughly equal. This gap implies that an environment is memory improvable: where performance gains in a domain are from an algorithm's ability to cope with partial observability as opposed to other factors. We introduce best-practice guidelines for empirically benchmarking reinforcement learning under partial observability, as well as the open-source library POBAX: Partially Observable Benchmarks in JAX. We characterize the types of partial observability present in various environments and select representative environments for our benchmark. These environments include localization and mapping, visual control, games, and more. Additionally, we show that these tasks are all memory improvable and require hard-to-learn memory functions, providing a concrete signal for partial observability research. This framework includes recommended hyperparameters as well as algorithm implementations for fast, out-of-the-box evaluation, as well as highly performant environments implemented in JAX for GPU-scalable experimentation.

## Contribution(s)

1. We investigate the efficacy of partially observable benchmarks in measuring an algorithm's ability to mitigate partial observability.
   **Context:** None
2. We introduce the memory improvability property: a partially observable benchmark is memory improvable if there is a gap between agents with more or less state information, all other factors roughly equal.
   **Context:** None
3. We categorize popular forms of partial observability, and present a list of representative environments that covers these categories.
   **Context:** This categorization does not cover all forms of partial observability.
4. We present the open-source POBAX benchmark: a suite of memory improvable environments designed to test an algorithm's ability to mitigate partial observability. POBAX is entirely implemented in JAX, allowing for fast and GPU-scalable experimentation.
   **Context:** While previous benchmarks exist for partial observability (Rajan et al., 2021; Morad et al., 2023; Osband et al., 2020), these works do not cover such breadth of environments.

# Rectifying Regression in Reinforcement Learning

**Alex Ayoub, David Szepesvári, Alireza Baktiari, Csaba Szepesvári, Dale Schuurmans**

**Keywords:** Value-based methods, Regression, Loss functions.

## Summary

This paper investigates the impact of the loss function in value-based methods for reinforcement learning through an analysis of underlying prediction objectives. We theoretically show that mean absolute error is a better prediction objective than the traditional mean squared error for controlling the learned policy's suboptimality gap. Furthermore, we present results that different loss functions are better aligned with these different regression objectives: binary and categorical cross-entropy losses with the mean absolute error and squared loss with the mean squared error. We then provide empirical evidence that algorithms minimizing these cross-entropy losses can outperform those based on the squared loss in linear reinforcement learning.

## Contribution(s)

1. We demonstrate certain cross entropy losses can accelerate convergence under certain structural assumptions, supported by negative results for the purely mean-focused squared loss.
   **Context:** We build upon a recent line of theoretical (Foster & Krishnamurthy, 2021; Ayoub et al., 2024; Wang et al., 2024) and empirical (Bellemare et al., 2017; Dabney et al., 2018; Farebrother et al., 2024) research showing that value learning with certain loss functions can yield faster convergence rates under specific structural assumptions, such as the optimal policy achieving the maximum possible value or having low variance returns. We complement these findings by providing lower bounds that link these convergence rates to the chosen regression objective—in this case mean absolute error and mean squared error.
2. We provide empirical results showing that value-based methods using log-loss (and its reparameterized multi-class variant) can outperform squared-loss methods in a linear batch reinforcement learning setting (inverted pendulum with Fourier features).
   **Context:** The work of Lyle et al. (2019) suggest that, in linear reinforcement learning, cross-entropy losses (e.g., binary or categorical) perform on par with squared loss and that their advantages appear primarily in deep reinforcement learning settings. However, our theoretical and empirical findings suggest a more subtle situation: in linear reinforcement learning, cross-entropy losses can outperform the canonical squared loss. Our experiments are limited to a single environment (inverted pendulum) with Fourier features.

# High-Confidence Policy Improvement from Human Feedback

**Hon Tik Tse, Philip S. Thomas, Scott Niekum**

**Keywords:** Reinforcement Learning from Human Feedback, High-Confidence Policy Improvement, Imitation Learning and Inverse Reinforcement Learning, Reinforcement Learning

## Summary

Reinforcement learning from human feedback (RLHF) aims to learn or fine-tune policies via human preference data when a ground-truth reward function is not known. However, many conventional RLHF methods provide no performance guarantees and have an unacceptably high probability of returning poorly performing policies. We propose Policy Optimization and Safety Test for Policy Improvement (POSTPI), an algorithm that provides high-confidence policy performance guarantees without direct knowledge of the ground-truth reward function, given only a preference dataset. The user of the algorithm may select any initial policy $\pi_{\text{init}}$ and confidence level $1 - \delta$, and POSTPI will ensure that the probability it returns a policy with performance worse than $\pi_{\text{init}}$ under the unobserved ground-truth reward function is at most $\delta$. We show theory as well as empirical results in the Safety Gymnasium suite that demonstrate that POSTPI reliably provides the desired guarantee.

## Contribution(s)

1. We formalize the problem of high-confidence policy improvement from human feedback (HCPI-HF).
   **Context:** Reinforcement learning from human feedback has been popular in recent years. However, the problem of performing high-confidence policy improvement from human preference data has not been formalized.
2. To address the HCPI-HF problem, we propose a novel algorithm Policy Optimization and Safety Test for Policy Improvement (POSTPI), and demonstrate both theoretically and empirically that POSTPI reliably provides the desired high-confidence policy improvement guarantee.
   **Context:** Many prior works in RLHF (Brown et al., 2019b; 2020; Javed et al., 2021; Hejna et al., 2024) provide no performance guarantees on the returned policy. While there exist some works that provide performance guarantees (Zhu et al., 2023; Chen et al., 2022; Xu et al., 2020; Pacchiano et al., 2023; Novoseller et al., 2020; Wang et al., 2023), different from these works, we focus specifically on the setting of improving with respect to a user-provided policy with high probability.
3. We propose a novel policy optimization objective that allows POSTPI to return a policy with high probability when the initial policy is sub-optimal, and derive the gradient of this objective.
   **Context:** Unlike PG-BROIL (Javed et al., 2021), which optimizes the conditional value-at-risk, we optimize the value-at-risk, and explicitly allow the objective to depend on the user-provided initial policy.
4. We propose a novel method for computing high-confidence policy performance bounds in the RLHF setting.
   **Context:** Unlike a prior approach (Brown et al., 2020), which only considers the uncertainty in the ground-truth reward function, our approach further considers the uncertainty in using a finite number of rollouts to estimate the expected value of a policy.

# Adaptive Reward Sharing to Enhance Learning in the Context of Multiagent Teams

**Kyle Tilbury, David Radke**

**Keywords:** Multiagent Reinforcement Learning, Meta-Reinforcement Learning, Coordination

## Summary

Real-world populations often include diverse social structures such as sub-groups or teams, creating heterogeneous incentives that complicate coordination. Therefore, autonomous agents must be able to adapt their individual incentives based on their surrounding population. To address this challenge, we introduce a decentralized multiagent reinforcement learning framework in which each individual agent learns to adapt both its behavior and its reward-sharing strategy within a defined social structure in mixed-motive environments. Inspired by meta-RL, each agent in our framework maintains two policies: a low-level behavioral policy and a high-level reward-sharing policy that updates its individual reward function, changing how agents distribute earned rewards and thereby shaping the incentives within the population. We demonstrate the viability of this self-tuning approach by showing how agent populations can learn to coordinate more effectively via the simultaneous adaptation of heterogeneous incentive configurations. This work is a step toward integrating learning agents into real-world scenarios with complex social structures and varying incentives.

## Contribution(s)

1. We introduce a multi-level framework enabling individual agents to not only learn behavior, but also adapt heterogeneous reward-sharing parameters in complex environments.
   **Context:** Recent work often assumes homogeneous or fixed reward-sharing schemes, limiting agents' ability to adapt to evolving social contexts (Durugkar et al., 2020; Radke et al., 2023a). Other approaches let agents learn to share or gift rewards but they either lack social structure or assume a cooperative global objective (Lupu & Precup, 2020; Yi et al., 2022). Our work, which considers mixed-motive environments with social structure, allows agents to learn and adapt to heterogeneous reward-sharing schemes, better capturing more diverse dynamics with social structures and varying incentives.
2. We show that our framework enhances population-level coordination, overcoming sub-optimal initialization and surpassing non-adaptive fully-cooperative baselines.
   **Context:** Using a standard but sub-optimal fully cooperative initialization, we demonstrate that adapting reward-sharing parameters enables agents to exceed the performance of non-adaptive fully-cooperative baselines by 34.2% and 20.3% (in mean population reward) across our two evaluation environments. Additionally, inspecting agent behaviors in one environment reveals that our adaptive agents consistently learn the best observed joint policy identified in prior work (Radke et al., 2023a).
3. We demonstrate that the heterogeneous reward-sharing parameterizations learned by our framework are highly effective when used to train new agent populations.
   **Context:** Using the heterogeneous reward-sharing schemes discovered during online tuning (i.e., while agents dynamically update their reward-sharing parameters during learning) as static parameterizations for newly instantiated agents yields further performance gains in both evaluation environments over online tuning. In one environment, these new populations surpass the best known configuration from prior work, achieving the highest observed reward while maintaining significantly greater equality. Discovering effective heterogeneous configurations through exhaustive or heuristic search of existing methods is onerous, our approach leverages reinforcement learning to autonomously learn effective solutions.

# MixUCB: Enhancing Safe Exploration in Contextual Bandits with Human Oversight

**Jinyan Su, Rohan Banerjee, Jiankai Sun, Wen Sun, Sarah Dean**

**Keywords:** Safe Exploration, human-in-the-loop contextual bandit

## Summary

The integration of AI into high-stakes decision-making domains demands safety and accountability. Traditional contextual bandit algorithms for online and adaptive decision-making must balance exploration and exploitation, posing significant risks when applied to critical environments where exploratory actions can lead to severe consequences. To address these challenges, we propose MixUCB, a flexible human-in-the-loop contextual bandit framework that enhances safe exploration by incorporating human expertise and oversight with machine automation. Based on the model's confidence and the associated risks, MixUCB intelligently determines when to seek human intervention. The reliance on human input gradually reduces as the system learns and gains confidence. Theoretically, we analyzed the regret and query complexity in order to rigorously answer the question of when to query. Empirically, we validate the effectiveness through extensive experiments on both synthetic and real-world datasets. Our findings underscore the importance of designing decision-making frameworks that are not only theoretically and technically sound, but also align with societal expectations of accountability and safety. Our experimental code is available at: https://github.com/sdean-group/MixUCB.

## Contribution(s)

1. We introduce **MixUCB**, a novel human-in-the-loop contextual bandit framework that dynamically determines when to seek human intervention based on uncertainty, enhancing safe exploration in high-stakes decision-making tasks. MixUCB is flexible in accepting various types of expert feedback.
   **Context:** Our approach unifies learning from experts (as in active learning, imitation learning, etc.) with learning from experience (as in reinforcement learning).
2. We provide a theoretical analysis of our framework, offering guarantees on regret and query complexity. This addresses the fundamental question of when to rely on expert input while balancing the cost and quality of the feedback.
   **Context:** While traditional online learning or bandit algorithms focus on fixed feedback settings, our analysis demonstrates MixUCB's adaptability to varying levels of expert involvement.
3. We demonstrate the practical effectiveness of MixUCB through experiments on both synthetic and real-world datasets, showing that the combination of human expertise and AI can outperform fully automated decision-making. We highlight the importance of designing AI systems that are not only technically sound but also emphasize safety, accountability, and human-centric decision-making.
   **Context:** Our experiments cover a range of feedback settings, showcasing MixUCB's ability to maintain high performance even when expert feedback is limited or noisy, for a domain-specific appropriate querying threshold.

# Efficient Morphology-Aware Policy Transfer to New Embodiments

**Michael Przystupa**[1,3,4], **Hongyao Tang**[3,4], **Martin Jagersand**[1], **Santiago Miret**[5], **Mariano Phielipp**[5], **Matthew E. Taylor**[1,2], **Glen Berseth**[3,4]

**Keywords:** Transfer Learning, Morphology-Aware Learning, Online Learning

## Summary

In this work, we investigate means of reducing the computation costs to finetune pre-trained morphology-aware policies to target morphologies with on-policy learning. Morphology-aware learning is a paradigm which attempts to learn several optimal policies across agent embodiments in a *single* neural network. A limitation of prior works have been focusing on end-to-end finetuning to adapt these policies to a target morphology. We address this gap by exploring parameter efficient techniques used successfully in other domains such as computer vision or natural language processing to specialize a policy. Our results suggest that using as few as 1% of total learnable parameters as the pre-trained model, we can achieve statistically significant performance improvements.

## Contribution(s)

1. We conduct an extensive series of experiments to compare the effects of parameter-efficient finetuning methods in the morphology-aware policy learning setting.
   **Context:** Prior works which include transfer learning experiments have generally focused on end-to-end finetuning or else at most consider low-rank adapter layers (LoRA), a form of delta weight learning, as part of their experiments (Octo Model Team, 2024). When LoRA has been used, experiments have only been conducted only in the *behavioral cloning* setting. This is a limitation in the literature because a wide variety of parameter-efficient techniques have been investigated in other fields such as prefix tuning in large language models (Li & Liang, 2021) and direct-finetuning in computer vision (Lee et al., 2023).
2. We are the first work to successfully learn policies using prefix tuning methods in the reinforcement learning settings.
   **Context:** Prefix tuning has been almost exclusively investigate in supervised learning settings such as natural language processing (Li & Liang, 2021), computer vision (Nie et al., 2023), or continual learning (Wang et al., 2022). The closest related to our work is Liu et al. (2024) who investigate prefix tuning techniques in the imitation learning setting and across *tasks* as opposed to agent morphology.
3. Our experiments reveal a number of trends in the morphology-aware policy setting. Generally we find that both input-adapter and prefix tuning methods converge to behaving similar to tuning the decoder head of the base model. Prefix tuning is particularly sensitive to hyperparameter choices where some configurations notably affect performance at the beginning of training and never recover. Generally, more parameters are always beneficial to improving policy performance in the tasks we considered.
   **Context:** Other such prescriptive research has been done in computer vision or language when investigating different PEFT techniques. The work of Lester et al. (2021) demonstrated the potential of prefix tuning over a number of factors including prompt initialization and number of prompt tokens. The work of Liu et al. (2022) highlights the benefits of injecting prompts in multiple layers in transformers. The work of Lee et al. (2023) suggests that intelligent layer different types of domain shifts in computer vision.

# Understanding Learned Representations and Action Collapse in Visual Reinforcement Learning

**Xi Chen, Zhihui Zhu, Andrew Perrault**

**Keywords:** Visual reinforcement learning, representation understanding.

## Summary

In contrast to deep learning models trained with supervised data, visual reinforcement learning (VRL) models learn to represent their environment implicitly via the process of seeking higher rewards. However, there has been little research on the specific representations VRL models learn. Using linear probing, we study the extent to which VRL models learn to linearly represent the ground truth vectorized state of an environment, on which layers these representations are most accessible, and how this relates to the reward achieved by the final model. We observe that poorly performing agents differ substantially from well-performing ones in the representation learned in their later MLP layers, but not their earlier CNN layers. When an agent is initialized by reusing the later layers of a poorly performing agent, the result is always poor. These poorly performing agents end up with no entropy in their actor network output, a phenomenon we call action collapse. Based on these observations, we propose a simple rule to prevent action collapse during training, leading to better performance on tasks with image observations with no additional computational cost.

## Contribution(s)

1. We present a case study showing how a VRL agent learns linear representations of the ground truth vectorized environment states using Orthogonal Matching Pursuit (OMP), i.e., linear probing with a sparsity constraint.
   **Context:** Linear probing has been widely used to study representations in other domains, but not in VRL.
2. In the CNN-to-MLP architecture we examine, the results of linear probing show that well- and poorly performing agents differ primarily in their later MLP layers, but not their earlier CNN layers.
   **Context:** This is counter to the intuition that the CNN layers are primarily responsible for representation learning.
3. The gap in the MLP but not the CNN layers between well-performing and poorly performing agents, revealed by linear probing results, is predictive of agent quality after retraining.
   **Context:** Linear probing has been questioned because it assumes the features are linearly accessible from learned representations.
4. We identify that the MLP layers of a poorly performing agent suffer from action collapse, a failure mode where all inputs in our experiments to an actor produce the same output.
   **Context:** Dormant and dead neurons have previously been observed in VRL (Xu et al., 2023), but action collapse is a more precise understanding of this failure mode.
5. By studying the metrics associated with action collapse, we show that it can be avoided with a simple rule, leading to zero poorly performing agents.
   **Context:** None

# Mitigating Suboptimality of Deterministic Policy Gradients in Complex Q-functions

**Ayush Jain, Norio Kosaka, Xinhu Li, Kyung-Min Kim, Erdem Bıyık, Joseph J. Lim**

**Keywords:** Deterministic Policy Gradients, Off-policy reinforcement learning

## Summary

In reinforcement learning, off-policy actor-critic methods such as DDPG and TD3 use deterministic policy gradients: the Q-function is learned from environment interaction data, while the actor seeks to maximize it via gradient ascent. We observe that in complex tasks—such as dexterous manipulation, restricted locomotion, and large discrete-action recommender systems—the Q-function exhibits multiple local optima, making naive gradient-based methods prone to getting stuck. To address this, we introduce Successive Actors for Value Optimization (SAVO), an architecture that (i) learns multiple actor networks, each conditioned on previously discovered actions, and (ii) employs a sequence of "surrogate" Q-landscapes that progressively truncate lower-value regions. This iterative scheme improves the global maximization of the Q-function while preserving the sample efficiency advantages of gradient-based updates. Experiments on restricted locomotion, dexterous manipulation, and recommender-system tasks demonstrate that SAVO outperforms single-actor methods as well as alternative multi-actor and sampling-based approaches.

## Contribution(s)

1. We propose a new multi-actor architecture that learns several policies in parallel and then selects the best action among them based on the current Q-function.
   **Context:** In deterministic policy gradient methods, a single actor frequently converges to local maxima of the Q-landscape. By training multiple actors and picking the highest-valued action, the final policy strictly improves over any single actor policy.
2. We introduce "successive surrogate" Q-functions that flatten out regions below previously discovered high-value actions, thus preventing actors from re-converging to known poor local optima.
   **Context:** Surrogate functions are created by lifting the Q-values in regions below an anchor action. This reduces the number of local maxima in the Q-landscape. We approximate these surrogates with neural networks to preserve gradient flow toward high-value regions without sacrificing expressiveness.
3. We demonstrate that our Successive Actors for Value Optimization (SAVO) method consistently yields higher returns in challenging tasks, including restricted continuous-control locomotion, dexterous manipulation, and large discrete-action recommender systems.
   **Context:** Standard TD3 or DDPG struggles in non-convex domains with many shallow local maxima, while evolutionary methods can be computationally expensive. Our approach combines the sample-efficiency of gradient-based learning with a mechanism to escape suboptimal local optima. Extensive ablations show that each element (multiple actors, surrogates, and conditioning on prior actions) contributes to performance gains.

# Leveraging priors on distribution functions for multi-arm bandits

**Sumit Vashishtha, Odalric-Ambrym Maillard**

**Keywords:** Bayesian nonparametric statistics, reinforcement learning, information theory

## Summary

We introduce Dirichlet Process Posterior Sampling (DPPS), a Bayesian non-parametric algorithm for multi-arm bandits based on Dirichlet Process (DP) priors. Like Thompson-sampling, DPPS is a probability-matching algorithm, i.e., it plays an arm based on its posterior-probability of being optimal. Instead of assuming a parametric class for the reward generating distribution of each arm, and then putting a prior on the parameters, in DPPS the reward generating distribution is directly modeled using DP priors. DPPS provides a principled approach to incorporate prior belief about the bandit environment, and in the noninformative limit of the DP priors (i.e. Bayesian Bootstrap), we recover Non Parametric Thompson Sampling (NPTS), a popular non-parametric bandit algorithm, as a special case of DPPS. We employ stick-breaking representation of the DP priors, and show excellent empirical performance of DPPS in challenging synthetic and real world bandit environments. Finally, using an information-theoretic analysis, we show non-asymptotic optimality of DPPS in the Bayesian regret setup.

## Contribution(s)

1. We introduce Dirichlet Process Posterior Sampling (DPPS) for multi arm bandits - a Bayesian nonparametric extension of Thompson sampling based on Dirichlet Processes that combines the strength of (Bayesian) bootstrap with a principled mechanism of incorporating and exploiting prior information.
   **Context:** Efficient performance of parametric Thompson sampling is limited to bandit environments wherein it's possible to have conjugate prior/posterior distributions. Besides, existing Bootstrap based algorithms cannot account for uncertainity that doesn't come from observed data (Osband et al., 2018)
2. We employ stick-breaking representation of the Dirichlet Process priors to perform numerical experiments with DPPS in both synthetic and real-world multi-arm bandit settings.
   **Context:** Improved performance of DPPS compared to parametric Thompson-sampling and UCB is made apparent in these simulations. Using a simple example, we also illustrate a proof-of-concept of the flexibility of DPPS in incorporating prior-knowledge about the bandit environment. Besides, Stick-Breaking implementation of DPPS provides a unified implementation for different bandit environments unlike parametric Thompson sampling whose implementation differ according to bandit environments and require careful tuning/approximations.
3. We extend the information theoretic analysis of Thompson sampling in Russo & Van Roy (2016) to a wider class of probability-matching algorithms that derive their posterior probability of optimal action using a valid Bayesian approach, and use this extension to establish $\sigma\sqrt{2\mathrm{TK}\log\mathrm{K}}$ non-asymptotic upper bound on the Bayesian regret of DPPS in bandit environments with $\sigma$ sub-Gaussian reward noise, where $\mathrm{T}$ is the time horizon, and $\mathrm{K}$ is the number of arms.
   **Context:** We are unaware of any Bootstrap based bandit algorithm that enjoys the order-optimal, $\sigma\sqrt{2\mathrm{TK}\log\mathrm{K}}$, non-asymptotic regret bound in the wide class of $\sigma$-sub-Gaussian bandit environments.

# ProtoCRL: Prototype-based Network for Continual Reinforcement Learning

**Michela Proietti, Peter R. Wurman, Peter Stone, Roberto Capobianco**

**Keywords:** Continual reinforcement learning, experience replay, event tables, prototype-based architecture, Gaussian mixture model, variational inference

## Summary

The purpose of continual reinforcement learning is to train an agent on a sequence of tasks such that it learns the ones that appear later in the sequence while retaining the ability to perform the tasks that appeared earlier. Experience replay is a popular method used to make the agent remember previous tasks, but its effectiveness strongly relies on the selection of experiences to store. Kompella et al. (2023) proposed organizing the experience replay buffer into partitions, each storing transitions leading to a rare but crucial event, such that these key experiences get revisited more often during training. However, the method is sensitive to the manual selection of event states. To address this issue, we introduce ProtoCRL, a prototype-based architecture leveraging a variational Gaussian mixture model to automatically discover effective event states and build the associated partitions in the experience replay buffer. The proposed approach is tested on a sequence of MiniGrid environments, demonstrating the agent's ability to adapt and learn new skills incrementally.

## Contribution(s)

1. This paper introduces ProtoCRL, a prototype-based architecture for continual reinforcement learning. ProtoCRL features a variational Gaussian mixture model to automatically identify effective event states and build the associated event tables, i.e., partitions within the experience replay buffer (ERB) storing transitions that lead to a particular event state.
   **Context:** Experience replay is a common strategy used in continual reinforcement learning (Liotet et al., 2022; Luo et al., 2023). Kompella et al. (2023) showed that partitioning the ERB into event tables increases sample efficiency and improves the agent's generalization performance. However, the method is sensitive to the manual selection of event states. ProtoCRL automatizes the construction of the ERB, making event tables suitable to applications in which the identification of event states is nontrivial.
2. The learned Gaussian mixture components practically serve as prototypical representations of an event state. By inspecting the assignments of the input experiences to the Gaussian mixture components, we show that ProtoCRL identifies meaningful event states that the agent needs to visit more often to remember previously learned tasks.
   **Context:** In the literature, prototypes have been used to either explain pre-trained black-box agents (Borzillo et al., 2023) or to improve the agents generalization performance of agents trained on single tasks (Liu et al., 2023). In this work, we leverage the learned prototypical representations to both guide experience replay and gain insights into what information is useful for the agents to maintain the ability to perform multiple tasks learned in sequence.
3. We show that ProtoCRL achieves comparable performance to manually defined event tables and even higher performance when reducing the ERB capacity.
   **Context:** We test ProtoCRL on a sequence of three MiniGrid environments (Chevalier-Boisvert et al., 2018), comparing its performance in terms of average return and forgetting to manually defined event tables and to ContinualDreamer (Kessler et al., 2023).

# Finer Behavioral Foundation Models via Auto-Regressive Features and Advantage Weighting

**Edoardo Cetin, Ahmed Touati, Yann Ollivier**

**Keywords:** Unsupervised RL, offline training, auto-regressive features, successor measures

## Summary

The forward-backward representation (FB) is a recently proposed framework (Touati et al., 2023; Touati & Ollivier, 2021) to train *behavior foundation models* (BFMs) that aim at providing zero-shot efficient policies for any new task specified in a given reinforcement learning (RL) environment, without training for each new task. Here we address two core limitations of FB model training.

First, FB, like all successor-feature-based methods, relies on a *linear* encoding of tasks: at test time, each new reward function is linearly projected onto a fixed set of pre-trained features. This limits expressivity as well as precision of the task representation. We break the linearity limitation by introducing *auto-regressive features* for FB, which let fine-grained task features depend on coarser-grained task information. This can represent arbitrary nonlinear task encodings, thus significantly increasing expressivity of the FB framework.

Second, it is well-known that training RL agents from offline datasets often requires specific techniques.We show that FB works well together with such offline RL techniques, by adapting techniques from (Nair et al., 2020b; Cetin et al., 2024) for FB. This is necessary to get non-flatlining performance in some datasets, such as DMC Humanoid.

As a result, we produce efficient FB BFMs for a number of new environments. Notably, in the D4RL locomotion benchmark, the generic FB agent matches the performance of standard single-task offline agents (IQL, XQL). In many setups, the offline techniques are needed to get any decent performance at all. The auto-regressive features have a positive but moderate impact, concentrated on tasks requiring spatial precision and task generalization beyond the behaviors represented in the trainset.

Together, these results establish that generic, reward-free FB BFMs can be competitive with single-task agents on standard benchmarks, while suggesting that expressivity of the BFM is not a key limiting factor in the environments tested.

## Contribution(s)

1. We overcome the linearity of reward representations in the Forward-Backward (FB) framework, without breaking the theoretical framework, thanks to *auto-regressive features* that let fine-grained task features depend on coarser-grained task information.
   **Context:** FB and Successor Features attempt to provide zero-shot RL adaptation to new rewards, but fundamentally rely on a linear reward encoding, which could restrict expressivity.
2. We show how to combine FB with offline RL techniques. We show this is necessary to get good performance in a number of environments and datasets, such as DMC Humanoid. We show that *generic, zero-shot* FB agents come close to the performance of recent *task-specific* agents on the D4RL benchmark.
   **Context:** Previous work has reported very poor FB performance in some situations (Park et al., 2024; Frans et al., 2024), due to omitting offline RL techniques.

# Pretraining Decision Transformers with Reward Prediction for In-Context Multi-task Structured Bandit Learning

**Subhojyoti Mukherjee, Josiah P. Hanna, Qiaomin Xie, Robert Nowak**

**Keywords:** Structured Bandit, Multi-task Learning, Decision Transformer

## Summary

We study learning to learn for the multi-task structured bandit problem where the goal is to learn a near-optimal algorithm that minimizes cumulative regret. The tasks share a common structure and an algorithm should exploit the shared structure to minimize the cumulative regret for an unseen but related test task. We use a transformer as a decision-making algorithm to learn this shared structure from data collected by a demonstrator on a set of training task instances. Our objective is to devise a training procedure such that the transformer will learn to outperform the demonstrator's learning algorithm on unseen test task instances. Prior work on pretraining decision transformers either requires privileged information like access to optimal arms or cannot outperform the demonstrator. Going beyond these approaches, we introduce a pre-training approach that trains a transformer network to learn a near-optimal policy in-context. This approach leverages the shared structure across tasks, does not require access to optimal actions, and can outperform the demonstrator. We validate these claims over a wide variety of structured bandit problems to show that our proposed solution is general and can quickly identify expected rewards on unseen test tasks to support effective exploration.

## Contribution(s)

1. We introduce a new pre-training and test time decision-making procedure that in-context learns the underlying reward structure for structured bandit settings, resulting in a near-optimal policy without access to privileged information even when training data comes from a sub-optimal demonstrator.
   **Context:** Previous works like DPT (Lee et al., 2023) required access to the optimal action per task, Algorithmic Distillation (AD) could not outperform the demonstrator, other works need to know the structure to perform optimally.
2. We show that our approach enables successful in-context learning across a diverse set of structured bandit settings where it matches the performance of existing algorithms that were developed with knowledge of the structure.
   **Context:** We evaluate our approach in linear,non-linear, bilinear, and latent bandit settings as well as bandit experiments based on real-life datasets and show that it lowers regret compared to DPT and AD while matching the near-optimal performance of specialized algorithms.
3. We show that our algorithm leverages the latent structure and conducts a two-phase exploration to minimize regret.
   **Context:** We analyze the exploration of the pretrained decision transformer in the simplified linear bandit setting where the optimal policy is well-understood. Previous works like DPT do not study the exploration conducted by such transformer algorithms. We introduce new actions both at train and test time. Since new actions are not shared across tasks now, the transformer algorithm fails to learn the latent structure as we scale up the number of new actions, thus indicating that it is relying on a discovered underlying structure. We observed in our experiments that our proposed algorithm implicitly conducts two-phase exploration, following the distribution of optimal action across training tasks and then switching to the most rewarding action for the task after observing a few in-context examples.

# Multi-task Representation Learning for Fixed Budget Pure-Exploration in Linear and Bilinear Bandits

**Subhojyoti Mukherjee, Qiaomin Xie, Robert Nowak**

**Keywords:** Multi-task learning, Pure Exploration, Linear Bandits, Bilinear Bandits

## Summary

We study fixed-budget pure exploration settings for multi-task representation learning (MTRL) in linear and bilinear bandits. In fixed budget MTRL linear bandit setting the goal is to find the optimal arm of each of the tasks with high probability within a pre-specified budget. Similarly, in a fixed budget MTRL bilinear setting the goal is to find the optimal left and right arms of each of the tasks with high precision within the budget. In both of these MTRL settings, the tasks share a common low-dimensional linear representation. Therefore, the goal is to leverage this underlying structure to expedite learning and identify the optimal arm(s) of each of the tasks with high precision.

We prove the first lower bound for the fixed-budget linear MTRL setting that takes into account the shared structure across the tasks. Motivated from the lower bound we propose the algorithm FB-DOE that uses a *double experimental design* approach to allocate samples optimally to the arms across the tasks, and thereby first learn the shared common representation and then identify the optimal arm(s) of each task. This is the first study on fixed-budget pure exploration of MTRL in linear and bilinear bandits. Our results show that learning the shared representation, jointly with allocating actions across the tasks following a double experimental design approach, achieves a smaller probability of error than solving the tasks independently.

## Contribution(s)

1. We formulate the first fixed-budget MTRL problem for the linear and bilinear bandit settings and establish the first lower bound for the fixed-budget MTRL linear bandit setting.
   **Context:** Previous work of MTRL setting studied fixed confidence linear (Du et al., 2023) and bilinear bandits (Mukherjee et al., 2023b). We establish the first lower bound for the fixed-budget MTRL in linear bandit setting and show that probability of error scales as $\widetilde{\Omega}(M\exp(-n\Delta^2/H_{2,\text{lin}}\log_2 k))$. Our bound contains the worst case hardness parameter $H_{2,\text{lin}}$ instead of the true hardness parameter $H_{1,\text{lin}}$. The work Du et al. (2023); Mukherjee et al. (2023b) provides no such lower bounds for the pure exploration MTRL setting.
2. We propose a double experimental design algorithm for fixed-budget MTRL linear bandits setting and prove a tight upper bound on the probability of error.
   **Context:** Our proposed algorithm for fixed-budget MTRL linear bandits has the probability of error scaling as $\widetilde{O}(M\exp(-n\Delta^2/H_{2,\text{lin}}\log_2 k))$. Therefore, the upper bound on the probability of error of our proposed algorithm matches the lower bound with respect to the parameters $k, d, M$, and worst case hardness $H_{2,\text{lin}}$. Previous work (Du et al., 2023) studied fixed confidence MTRL linear bandit setting.
3. We also extend our work to fixed-budget bilinear bandit settings and again propose a double experimental design algorithm.
   **Context:** Our proposed algorithm achieves a probability of error that scales as $\widetilde{O}(M(\exp(-n\Delta^2)/H_{2,\text{bilin}}\log_2(k_1+k_2)r)$. Previous work (Mukherjee et al., 2023b) studied fixed confidence MTRL bilinear bandit setting. We show the first upper bound on the probability of error in bilinear setting that has the worst case hardness parameter $H_{2,\text{bilin}}$ in the bound.

# Offline Reinforcement Learning with Domain-Unlabeled Data

**Soichiro Nishimori** **Xin-Qiang Cai**
**Johannes Ackermann** **Masashi Sugiyama**

**Keywords:** Offline Reinforcement Learning, Domain-Unlabeled Data, Positive-Unlabeled Learning, Weakly-supervised Learning

## Summary

Offline reinforcement learning (RL) is vital in areas where active data collection is expensive or infeasible, such as robotics or healthcare. In the real world, offline datasets often involve multiple "domains" that share the same state and action spaces but have distinct dynamics, and only a small fraction of samples are clearly labeled as belonging to the target domain we are interested in. For example, in robotics, precise system identification may only have been performed for part of the deployments. To address this challenge, we consider Positive-Unlabeled Offline RL (PUORL), a novel offline RL setting in which we have a small amount of labeled target-domain data and a large amount of domain-unlabeled data from multiple domains, including the target domain. For PUORL, we propose a plug-and-play approach that leverages positive-unlabeled (PU) learning to train a domain classifier. The classifier then extracts target-domain samples from the domain-unlabeled data, augmenting the scarce target-domain data. Empirical results on a modified version of the D4RL benchmark demonstrate the effectiveness of our method: even when only $1\%$–$3\%$ of the dataset is domain-labeled, our approach accurately identifies target-domain samples and achieves high performance, even under substantial dynamics shift. Our plug-and-play algorithm seamlessly integrates PU learning with existing offline RL pipelines, enabling effective multi-domain data utilization in scenarios where comprehensive domain labeling is prohibitive.

## Contribution(s)

1. We introduce Positive-Unlabeled Offline RL (PUORL), a novel offline RL setting with a small amount of data from a target domain and a large dataset containing data from multiple domains without domain labels. The goal is to learn a policy for the target domain.
   **Context:** Existing cross-domain offline RL methods (Liu et al., 2022; 2023; Wen et al., 2024) assume knowledge of the original domain of each transition, which is not accessible in our setting.
2. We propose a method that uses positive-unlabeled (PU) learning to filter the target-domain data from domain-unlabeled data.
   **Context:** Our approach uses PU learning (Li & Liu, 2003; Kiryo et al., 2017) to classify domain-unlabeled samples as "positive" (target) or "negative" (other). We then augment the labeled target-domain dataset with the domain-unlabeled samples predicted to be positive. This filtering can be integrated with value-based offline RL algorithms.
3. We empirically demonstrate that our PU-based method accurately filters domain-unlabeled data and achieves high performance in a modified version of D4RL.
   **Context:** We tested our approach on a modified D4RL benchmark (Fu et al., 2020), where only $1\%$–$3\%$ of samples contain domain labels, and the rest are domain-unlabeled, drawn from both the target and other domains with different dynamics. Even with this limited labeling, our method closely matches an oracle baseline (which has access to all target-domain data) and overall achieves higher average returns than the other baselines, even under substantial dynamics mismatch.

# Multi-Agent Reinforcement Learning for Inverse Design in Photonic Integrated Circuits

**Yannik Mahlau, Maximilian Schier, Christoph Reinders, Frederik Schubert, Marco Bügling, Bodo Rosenhahn**

**Keywords:** Photonic Integrated Circuits, MARL, Discrete Optimization, Optical Computing

## Summary

Inverse design of photonic integrated circuits (PICs) has traditionally relied on gradient-based optimization. However, this approach is prone to end up in local minima, which results in suboptimal design functionality. As interest in PICs increases due to their potential for addressing modern hardware demands through optical computing, more adaptive optimization algorithms are needed. We present a reinforcement learning (RL) environment as well as multi-agent RL algorithms for the design of PICs. By discretizing the design space into a grid, we formulate the design task as an optimization problem with thousands of binary variables. We consider multiple two- and three-dimensional design tasks that represent PIC components for an optical computing system. By decomposing the design space into thousands of individual agents, our algorithms are able to optimize designs with only a few thousand environment samples. They outperform previous state-of-the-art gradient-based optimization in both two- and three-dimensional design tasks. Our work may also serve as a benchmark for further exploration of sample-efficient RL for inverse design in photonics.

## Contribution(s)

1. We introduce the design of photonic integrated circuit components as a discrete optimization problem, which we implement as a multi-agent reinforcement learning (MARL) environment. This bandit-like MARL environment tests the interaction of multiple thousand agents with very few samples.
   **Context:** Photonic integrated circuits enable optical computing, which is a new field for fast and energy efficient hardware accelerators (McMahon, 2023). Previous research in MARL mostly focused on a handful of agents using millions of training samples to learn an environment with many states (Rutherford et al., 2023). In contrast, we introduce a bandit setting with a single environment state, but multiple thousands of agents using only 10000 training samples. The sample efficiency is important because electromagnetic simulations for environment steps are time-consuming (Mahlau et al., 2025).
2. To solve the challenges of our new environment, we develop two multi-agent reinforcement learning algorithms. They are based on proximal policy optimization (Schulman et al., 2017) and an actor-critic approach with stochastic policies similar to the soft-actor-critic (Haarnoja et al., 2018). In extensive experiments, we show that our algorithms outperform previous state-of-the-art.
   **Context:** Inverse design in photonics has previously almost exclusively been performed using gradient-based optimization (Schubert et al., 2025), which can quickly find a decent solution, but its susceptibility to getting stuck in local minima during optimization impedes performance.
3. We publish the reinforcement learning environment and training algorithms as open-source.
   **Context:** We hope to facilitate reproducibility and further research.

# Syllabus: Portable Curricula for Reinforcement Learning Agents

**Ryan Sullivan, Ryan Pégoud, Ameen Ur Rehman, Xinchen Yang, Junyun Huang, Aayush Verma, Nistha Mitra, John P. Dickerson**

**Keywords:** Curriculum Learning, Unsupervised Environment Design, Open-Endedness

## Summary

Curriculum learning has been a quiet, yet crucial component of many high-profile successes of reinforcement learning. Despite this, it is still a niche topic that is not directly supported by any of the major reinforcement learning libraries. These methods can improve the capabilities and generalization of RL agents, but often require complex changes to training code, limiting their impact on the field. We introduce Syllabus, a portable curriculum learning library, as a solution to this problem. Syllabus provides a universal API for curriculum learning, modular implementations of popular automatic curriculum learning methods, and infrastructure that allows them to be easily integrated with asynchronous training code in nearly any RL library. Syllabus provides a minimal API for core curriculum learning components, making it easier to design new algorithms and adapt existing ones to new environments. We demonstrate this by evaluating the algorithms in Syllabus on several new environments, each using agents written in a different RL library. We present the first examples of automatic curriculum learning in NetHack and Neural MMO, two of the most challenging RL benchmarks, and find evidence that existing methods do not easily transfer to complex new environments.

## Contribution(s)

1. This paper introduces Syllabus, a library of portable curriculum learning algorithms and infrastructure for synchronizing curricula across reinforcement learning environments running in separate processes. Syllabus includes portable implementations of several popular automatic curriculum learning algorithms and tools for manually designing curricula.
   **Context:** There are open-source curriculum learning libraries (Jiang et al., 2022; 2023; Dharna et al., 2022; Coward et al., 2024), but they build curriculum logic into the RL training code, making it difficult to extend methods and apply them to new environments. Syllabus is the first portable infrastructure for curriculum learning.
2. We evaluate tuned curriculum learning baselines in 4 environments including 2 which have not previously been explored in the context of curriculum learning. These baselines provide a solid foundation for future curriculum learning research.
   **Context:** We implement baselines from Jiang et al. (2021b), Kanitscheider et al. (2021), Zhang et al. (2023), and Rutherford et al. (2024) and apply all four algorithms to baseline environments used in these works. Some of these evaluations are reproductions of the experiments in those papers, but most are novel.
3. Our experiments demonstrate that popular curriculum learning methods are far less effective outside of the environments in which they were originally developed, and that more advanced methods may be necessary in complex environments.
   **Context:** Previous work successfully used automatic curricula over level seeds to train agents in procedurally generated environments. We show that these curricula over level seeds are ineffective in new or complex environments.

# Exploration-Free Reinforcement Learning with Linear Function Approximation

**Luca Civitavecchia, Matteo Papini**

**Keywords:** exploration-free, linear function approximation, no-regret.

## Summary

In the context of Markov Decision Processes (MDPs) with linear Bellman completeness, a generalization of linear MDPs, we reconsider the learning capabilities of a *greedy* algorithm. The motivation is that, when exploration is costly or dangerous, an exploration-free approach may be preferable to optimistic or randomized solutions. We show that, under a condition of sufficient diversity in the feature distribution, Least-Squares Value Iteration (LSVI) can achieve sublinear regret. Specifically, we show that the expected cumulative regret is at most $\tilde{\mathcal{O}}(H^3\sqrt{dK/\lambda_0})$, where $K$ is the number of episodes, $H$ is the task horizon, $d$ is the dimension of the feature map and $\lambda_0$ is a measure of feature diversity. We empirically validate our theoretical findings on synthetic linear MDPs. Our analysis is a first step towards exploration-free reinforcement learning in MDPs with large state spaces.

## Contribution(s)

1. The definition of a new diversity condition for linear MDPs.
   **Context:** Inspired from prior work of Bastani et al. (2021) and Kannan et al. (2018).
2. Proved that a greedy algorithm (LSVI) achieves sublinear cumulative regret with high probability when the here defined diversity condition is satisfied.
   **Context:** Proof built upon the related work on linear contextual bandit of Bastani et al. (2021).
3. Proved that a greedy algorithm (LSVI) achieves sublinear cumulative regret with high probability when the here defined diversity condition is satisfied, under a misspecified setting.
   **Context:** Proof built upon the related work on approximately linear MDPs of Zanette et al. (2020).

# SPEQ: Offline Stabilization Phases for Efficient Q-Learning in High Update-To-Data Ratio Reinforcement Learning

**Carlo Romeo*, Girolamo Macaluso*, Alessandro Sestini, Andrew D. Bagdanov**

**Keywords:** Reinforcement Learning, UTD, computational efficiency

## Summary

Reinforcement learning (RL) algorithms that employ high update-to-data (UTD) ratios have demonstrated significant improvements in sample efficiency by performing multiple updates per environment interaction. However, this strategy comes at a considerable computational cost that can render it impractical for large-scale or real-world applications where efficiency is paramount. In this work we propose Offline Stabilization Phases for Efficient Q-Learning (SPEQ), a novel RL algorithm that interleaves low-UTD online training with periodic offline stabilization phases. During these phases, Q-functions are fine-tuned with very high UTD ratios while keeping the replay buffer fixed, reducing redundant gradient updates on suboptimal data. To mitigate the overestimation bias problem due to the multiple and consecutive updates, SPEQ implements dropout regularization only for critics. This approach improves computational efficiency without compromising learning effectiveness. Empirical results on the MuJoCo continuous control benchmark demonstrate that SPEQ significantly reduces computational overhead while achieving performance comparable to state-of-the-art high UTD ratio methods.

## Contribution(s)

1. We propose **SPEQ**, a novel reinforcement learning algorithm that integrates periodic offline stabilization phases to balance computational and sample efficiency.
   **Context:** Our method contrasts with traditional high UTD ratio approaches by strategically concentrating computational resources in stabilization phases rather than uniformly distributing Q-function updates.
2. We empirically demonstrate that SPEQ requires from **40%** to **99%** fewer gradient updates and from **27%** to **78%** less training time compared to state-of-the-art, high UTD ratio reinforcement learning methods while maintaining competitive performance.
   **Context:** This highlights the computational advantages of structured training schedules over conventional high UTD ratio strategies.
3. We show that SPEQ outperforms simple reductions in UTD ratio, demonstrating that periodic stabilization phases provide a **more effective way** to optimize learning efficiency.
   **Context:** This distinction is crucial for designing more scalable and efficient reinforcement learning algorithms.

# Understanding Behavioral Metric Learning:

## A Large-Scale Study on Distracting Reinforcement Learning Environments

**Ziyan "Ray" Luo, Tianwei Ni, Pierre-Luc Bacon, Doina Precup, Xujie Si**

**Keywords:** behavioral metrics, bisimulation metrics, representation learning, evaluation

## Summary

A key approach to state abstraction is approximating *behavioral metrics* (notably, bisimulation metrics) in the observation space, and embedding these learned distances in the representation space. While promising for robustness to task-irrelevant noise, as shown in prior work, accurately estimating these metrics remains challenging, requiring various design choices that create gaps between theory and practice. Prior evaluations focus mainly on final returns, leaving the quality of learned metrics and the source of performance gains unclear. To systematically assess how metric learning works in deep reinforcement learning (RL), we evaluate five recent approaches, unified conceptually as isometric embeddings with varying design choices. We benchmark them with baselines across 20 state-based and 14 pixel-based tasks, spanning *370 task configurations* with diverse noise settings. Beyond final returns, we introduce the evaluation of a *denoising factor* to quantify the encoder's ability to filter distractions. To further isolate the effect of metric learning, we propose and evaluate an *isolated metric estimation* setting, in which the encoder is influenced solely by the metric loss. Finally, we release an open-source, modular codebase to improve reproducibility and support future research on metric learning in deep RL.

## Contribution(s)

1. We analyze five recent metric learning approaches under the isometric embedding framework to identify key design choices.
   **Context:** Metric learning methods often diverge significantly between theory and implementation.
2. We introduce the denoising factor to quantify an encoder's ability to filter distractions.
   **Context:** Metric learning is often motivated by denoising ability but is rarely evaluated directly, with prior work relying mainly on qualitative analysis (Zhang et al., 2020).
3. We benchmark five metric learning approaches across diverse distracting domains and find that common benchmarks add little difficulty to clean tasks, while certain noise settings remain challenging. We adopt most hyperparameters provided in their codebases, as we believe they are well-tuned for their benchmarks that share the same denoised states as ours.
   **Context:** Prior work primarily uses IID Gaussian noise with varied dimensions (Ni et al., 2024) and grayscale video backgrounds (Zhang et al., 2020).
4. Through ablation studies, we identify layer normalization and self-prediction loss as key design choices across all methods.
   **Context:** Prior work in metric learning does not isolate the effect of self-prediction loss and only shows the benefits of normalization in specific methods (Zang et al., 2022).
5. We show that the benefits of metric learning diminish in both return and denoising factor when key design choices are incorporated into the baseline.
   **Context:** Prior work does not report this limitation of metric learning.

# Collaboration Promotes Group Resilience in Multi-Agent RL

**Ilai Shraga, Guy Azran, Matthias Gerstgrasser, Ofir Abu, Jeffrey S. Rosenschein, Sarah Keren**

**Keywords:** Multi-Agent Reinforcement Learning, Group Resilience, Collaboration, Deep Reinforcement Learning.

## Summary

To effectively operate in various dynamic scenarios, RL agents must be resilient to unexpected changes in their environment. Previous work on this form of resilience has focused on single-agent settings. In this work, we introduce and formalize a multi-agent variant of resilience, which we term *group resilience*. We further hypothesize that collaboration with other agents is key to achieving group resilience; collaborating agents adapt better to environmental perturbations in multi-agent reinforcement learning (MARL) settings. We test our hypothesis empirically by evaluating different collaboration protocols and examining their effect on group resilience. Our experiments show that all the examined collaborative approaches achieve higher group resilience than their non-collaborative counterparts.

## Contribution(s)

1. We introduce a novel definition of group resilience and formalize this notion, which corresponds to the group's ability to adapt to unexpected changes.
   **Context:** Prior work primarily focused on resilience in single-agent settings or adversarial MARL scenarios without a unified resilience measure.
2. We demonstrate that collaboration promotes group resilience, providing empirical evidence across multiple MARL benchmarks.
   **Context:** Since prior work has mostly explored single-agent or adversarial settings, collaboration and its effects have not been considered.

# Value Bonuses using Ensemble Errors for Exploration in Reinforcement Learning

**Abdul Wahab, Raksha Kumaraswamy, Martha White**

**Keywords:** Reinforcement Learning, Exploration, Value bonuses, Ensembles, Uncertainty estimates

## Summary

Optimistic value estimation can be useful to direct exploration and improve sample efficiency in Reinforcement Learning. Despite many such methods in literature, simpler, undirected approaches like $\epsilon$-greedy still continue to be widely used. One potential reason for this is that many existing methods can be onerous to use as they may not be compatible with the base learning algorithm, or can be hard-to-use as many design choices need to be made to make them effective in practice. This paper proposes a simple approach to address these limitations. Building on ideas that utilize an ensemble for optimistic value estimation, this work proposes an algorithm called Value Bonuses using Ensemble Errors (VBE) that is easy to use and compatible with any base Reinforcement Learning algorithm, with a small additional computational foot-print. VBE's similarity and difference to existing approaches is discussed, and the algorithm is evaluated extensively. Our code is available at: https://github.com/mirzaabdulwahab1612/VBE

## Contribution(s)

1. Proposes a new approach for estimating Value Bonuses using Ensemble Errors that allows for first-visit optimism and deep exploration.
   **Context:** Many prior works utilize the idea of *value bonuses* to estimate optimistic values for exploration (Osband et al., 2019). Typically, many estimate an additional value function that propagates *reward bonuses* in order to estimate the value bonus (Burda et al., 2019). This work proposes a variant of value bonuses that does not rely on propagating additional reward bonuses; this allows for desirable features like first-visit optimism. We show our framework allows for Optimistic Initial Values with high probability. Additionally, the value bonuses have a similar timescale of learning as the main value function, therefore, potentially allowing for deep exploration.
2. Provides insight into how our proposed value bonuses are similar to and different from some relevant widely-used approaches.
   **Context:** In specific, we contrast how the proposed bonuses capture MDP-specific properties like transition dynamics, unlike those of RND (Burda et al., 2019). Additionally, we highlight the similarity between quantities estimated by the proposed algorithm and BDQN (Osband et al., 2019), despite the difference that BDQN utilizes a Thompson sampling approach to induce optimism.
3. Empirically evaluate the utility of the proposed value bonuses for inducing exploration in classic-control problems, and demonstrate scalability through Atari.
   **Context:** We demonstrate how existing methods lack first-visit optimism in a controlled setting designed to test state coverage. We show that the proposed algorithm can extend to more complex environments like Atari without design choices that alter the underlying algorithm.

# Gaussian Process Q-Learning for Finite-Horizon Markov Decision Processes

**Maximilian Bloor, Tom Savage, Calvin Tsay, Ehecatl Antonio Del Rio Chanona, Max Mowbray**

**Keywords:** Finite Horizon Markov Decision Processes, Gaussian Process, Q-learning

## Summary

Many real-world control and optimization problems require making decisions over a finite time horizon to maximize performance. This paper proposes a reinforcement learning framework that approximately solves the finite-horizon Markov Decision Process (MDP) by combining Gaussian Processes (GPs) with Q-learning. The method addresses two key challenges: the tractability of exact dynamic programming in continuous state-control spaces, and the need for sample-efficient state-action value function approximation in systems where data collection is expensive. Using GPs and backward induction, we construct state-action value function approximations that enable efficient policy learning with limited data. To handle the computational burden of GPs as data accumulate across iterations, we propose a subset selection mechanism that uses M-determinantal point processes to draw diverse, high-performing subsets. The proposed method is evaluated on a linear quadratic regulator problem and online optimization of a non-isothermal semi-batch reactor. Improved learning efficiency is shown relative to the use of Deep Q-networks and exact GPs built with all available data.

## Contribution(s)

1. The paper presents a framework for learning Gaussian process state-action value function approximations using Q-learning for deterministic finite horizon Markov Decision Processes.
   **Context:** Prior work has explored the use of Gaussian process models within the infinite horizon context but has shown no principled mechanism to handle increasing dataset size (Engel et al., 2005; Grande et al., 2014; Chowdhary et al., 2014).
2. A subset selection strategy is proposed to ensure the online computational tractability of control inference using M-determinantal point processes to build a GP approximation of the state-action value function, that balances global coverage with local accurate modeling in highly performing regions of the state-control space (Kulesza et al., 2012; Moss et al., 2023)
   **Context:** Previous works take the approach of building variational approximations globally to the exact GP state-action value function approximation (Grande et al., 2014).

# On the Effect of Regularization in Policy Mirror Descent

**Jan Felix Kleuker, Aske Plaat, Thomas Moerland**

**Keywords:** Policy Mirror Descent, Regularization, Reinforcement Learning

## Summary

Policy Mirror Descent (PMD) has emerged as a unifying framework in reinforcement learning (RL) by linking policy gradient methods with a first-order optimization method known as mirror descent. At its core, PMD incorporates two key regularization components: (i) a distance term that enforces a trust region for stable policy updates and (ii) an MDP regularizer that augments the reward function to promote structure and robustness. While PMD has been extensively studied in theory, empirical investigations remain scarce. This work provides a large-scale empirical analysis of the interplay between these two regularization techniques, running over 500k training seeds on small RL environments.

## Contribution(s)

(i) We conduct an empirical analysis of the interaction between the regularization components in PMD, running over 500k training seeds and providing insights into the fragility of these algorithms to regularization temperatures.
**Context:** Empirical studies on PMD-based RL algorithms are limited.

(ii) We examined the impact of temperature scales across different reward scales and identified simple heuristics to aid in tuning these temperatures. Additionally, we evaluated dynamic adaptation strategies for temperature parameters, indicating that maintaining constant values is often more effective than adapting them.
**Context:** RL algorithms are sensitive to hyperparameter tuning, making the selection of appropriate temperatures challenging.

(iii) We examine the effects of different regularization combinations on robustness to temperature tuning and performance, indicating a notable impact.
**Context:** To the best of our knowledge, no existing study examines the effects of the interplay between these two components for different choices, leaving a gap in the literature.

# Concept-Based Off-Policy Evaluation

**Ritam Majumdar, Jack Teversham, Sonali Parbhoo**

**Keywords:** Off Policy Evalutaion, Interpretability, Concept Bottleneck Models, Reliable OPE

## Summary

Evaluating off-policy decisions using batch data is challenging because of limited sample sizes which lead to high variance. Identifying and addressing the sources of this variance is crucial to improve off-policy evaluation in practice. Recent research on Concept Bottleneck Models (CBMs) shows that using human-explainable concepts can improve predictions and provide additional context for understanding decisions. In this paper, we propose incorporating an analogous notion of concepts into OPE to provide additional context that may help us identify specific areas where variance is high. We introduce a family of new concept-based OPE estimators and show that these estimators have two key properties when the concepts are known in advance: they remain unbiased whilst reducing variance of overall estimates. Since real-world applications often lack predefined concepts, we further develop an end-to-end algorithm to learn interpretable, concise, and diverse concepts optimized for variance reduction in OPE. Our experiments on synthetic and real-world datasets show that both known and learnt concept-based estimators significantly improve OPE performance. Crucially, our concept-based estimators offer two advantages over existing OPE methods. First, they are easily interpretable. Second, they allow us to isolate specific concepts contributing to variance. Upon performing targeted interventions on these concepts, we can further enhance the quality of OPE estimators.

## Contribution(s)

1. We introduce a new family of IS estimators based on interpretable concepts. [Section 3]
   **Context:** Previous works perform IS in the state representations, we explicitly define what is a concept representation and tie the original definition of IS under concepts.
2. We derive theoretical conditions ensuring lower variance compared to existing IS estimators. [Section 4]
   **Context:** We compare the variance of the Concept-OPE estimators with traditional IS/PDIS and MIS estimators and devise conditions under which the variance is reduced.
3. We propose an end-to-end algorithm for optimizing parameterized concepts when concepts are unknown, using OPE characteristics like variance. [Section 5]
   **Context:** Under real-world scenarios, the concepts are typically unknown or hard to define, which adds to the complexity of performing OPE. In this section, we propose a novel algorithm which learns concepts that satisfy the desidarata: Explainability, Conciseness, Diversity while optimizing for variance.
4. We show, through synthetic and real experiments, that our estimators for both known and unknown concepts outperform existing ones. [Sections 4,5]
   **Context:** None
5. We interpret the learned concepts to explain OPE characteristics and suggest intervention strategies to further improve OPE estimates. [Section 6]
   **Context:** Interventions have been typically studied in the context of improving the CBM performance in a supervised learning regime, we instead use interpretations to explain where a concept-OPE estimator has high variance and intervene to reduce variance.

# Investigating the Utility of Mirror Descent in Off-policy Actor-Critic

**Samuel Neumann, Jiamin He, Adam White, Martha White**

**Keywords:** mirror descent, off-policy, actor-critic, Soft Actor-Critic, Greedy Actor-Critic, Maximum A-Posteriori Policy Optimization

## Summary

Many policy gradient methods prevent drastic changes to policies during learning. This is commonly achieved through a Kullback-Leibler (KL) divergence term. Recent work has established a theoretical connection between this heuristic and Mirror Descent (MD), offering insight into the empirical successes of existing policy gradient and actor-critic algorithms. This insight has further motivated the development of novel algorithms that better adhere to the principles of MD, alongside a growing body of theoretical research on policy mirror descent. In this study, we examine the empirical feasibility of MD-based policy updates in off-policy actor-critic. Specifically, we introduce principled MD adaptations of three widely used actor-critic algorithms and systematically evaluate their empirical effectiveness. Our findings indicate that, while MD-style policy updates are not significantly advantageous over conventional approaches to actor-critic, they can somewhat mitigate sensitivity to step size selection with widely used deep-learning optimizers.

## Contribution(s)

1. We derive novel Mirror Descent variants of Soft Actor-Critic (SAC), Greedy Actor-Critic (GreedyAC), and Maximum A-Posteriori Policy Optimization (MPO) based on the Functional Mirror Descent (FMD) perspective.
   **Context:** A growing body of work on Policy Mirror Descent (PMD) has theoretically motivated the benefits of using Mirror Descent (MD) to update policies (Xiao, 2022; Johnson et al., 2023; Fatkhullin & He, 2024; Vieillard et al., 2020a; Lan, 2023). Much of this theory is for the tabular setting with exact policies. Recent work went one step further and re-derived several policy gradient algorithms for function approximation by introducing a functional MD (FMD) perspective (Vaswani et al., 2022). Such a perspective has yet to be brought to off-policy actor-critic methods that use approximate action-values and alternative losses for the actor. We are not claiming to have introduced the FMD perspective, nor that our derivations are complex, but they produce new algorithms.
2. We show these new MD variants exhibit no significant performance advantage over SAC, GreedyAC, and MPO across a variety of small problems and MuJoCo tasks.
   **Context:** It is possible there could be a difference in different environments.
3. We find that these MD algorithms provide (1) minor improvement in sensitivity to actor step size, (2) no improvement in sensitivity to entropy regularization parameter, and (3) no improvement with increasing replay ratio for actor updates; even though all three potential benefits are suggested by the theory.
   **Context:** Recent work suggests that policy gradient algorithms often encounter cliffs in the gradient direction, limiting step size magnitudes and explaining sensitivity (Jordan et al., 2024; Sullivan et al., 2022). Since MD updates account for policy-space distances, they should be more robust to step sizes. The KL in MD updates may already prevent policy collapse by regulating policy changes, hence these algorithms should be less sensitive to entropy regularization. Finally, MD updates prevent the algorithm from changing the policy too much, in probability space, and thus the amount of replay per step can be increased. One actor update corresponds to an approximate MD step; increasing the number of actor updates better approximates an exact MD step, which theoretically should perform better.

# Hybrid Classical/RL Local Planner for Ground Robot Navigation

**Vishnu Dutt Sharma, Jeongran Lee, Matthew Andrews, Ilija Hadžić**

**Keywords:** Mobile robot navigation, Local path planning, Hybrid planning

## Summary

Local planning is an optimization process within a mobile robot navigation stack that searches for the best velocity vector, given the robot and environment state. Depending on how the optimization criteria and constraints are defined, some planners may be better than others in specific situations. We consider two conceptually different planners. The first planner explores the velocity space in real-time and has the robot's dynamic model. It has superior path-tracking and motion smoothness performance. The second planner was trained using reinforcement learning methods to avoid obstacles. It is better at avoiding dynamic obstacles, but at the expense of motion smoothness. We propose a simple, yet effective, meta-reasoning approach that takes advantage of both approaches by switching between planners based on the surroundings. We demonstrate the superiority of our hybrid planner, both qualitatively and quantitatively, over individual planners on a live robot in different scenarios, achieving an improvement of **26%** in the navigation time.

## Contribution(s)

1. This paper present a hybrid local planner for ground robots that uses a classical planner when the immediate environment is simple, and an RL-based planner for more complex local environments.
   **Context:** There have been many recent efforts that apply RL to ground robot local planners. To the best of our knowledge, these use RL all the time, which we believe is excessive for simple environments where classical planners work well.
2. A key contribution is a simpler criterion that decides whether the classical planner or the RL-based planner should be used.
   **Context:** Training one RL planner that operates well in every environment is a difficult and impractical task. Specifically, the motion smoothness can be improved by adding robot dynamics to the reward function at the expense of making the training more difficult. Alternatively different (more specialized) models can be used depending on the situation (classical model that knows robot dynamics vs. learned RL model in our case). We show that a simple decision criterion is sufficient to achieve the benefit from both models and that training another network to implement this meta-policy is unnecessary.
3. We integrate our hybrid local planner into a full ROS stack and implement on physical robots.
   **Context:** Direct deployment of RL-based planners to real-world may not result in efficient operation and may even require fine-tuning in real-world. By implementing and testing our system on physical robots, we show that the proposed solution is applicable in practice.
4. We demonstrate via extensive simulations that our hybrid local planner achieves the "best of both worlds", in that it balances between travel time for simple environments and collision avoidance for more complex dynamic environments.
   **Context:** Previously reported work on hybrid planning focus mainly on collision avoidance for "social navigation".

# How Should We Meta-Learn Reinforcement Learning Algorithms?

**Alexander D. Goldie, Zilin Wang, Jaron Cohen, Jakob N. Foerster, Shimon Whiteson**

**Keywords:** Meta-Reinforcement Learning, Algorithm Discovery.

## Summary

The process of meta-learning algorithms from data, instead of relying on manual design, is growing in popularity as a paradigm for improving the performance of machine learning systems. Meta-learning shows particular promise for reinforcement learning (RL), where algorithms are often adapted from supervised or unsupervised learning despite their suboptimality for RL. However, until now there has been a severe lack of comparison between different meta-learning algorithms, such as using evolution to optimise over black-box functions or LLMs to propose code. In this paper, we carry out this empirical comparison of the different approaches when applied to a range of meta-learned algorithms which target different parts of the RL pipeline. In addition to meta-train and meta-test performance, we also investigate factors including the interpretability, sample cost and train time for each meta-learning algorithm. Based on these findings, we propose several guidelines for meta-learning new RL algorithms which will help ensure that future learned algorithms are as performant as possible.

## Contribution(s)

1. We provide an empirical study of a number of different meta-learning algorithms when applied to a range of meta-learned algorithms for online reinforcement learning. This study considers the trade-off of meta-train and meta-test performance of learned algorithms in addition to how sample-efficient, time-consuming and interpretable different meta-learning algorithms are.
   **Context:** Prior work has introduced a number of different meta-learning algorithms, such as using evolution to optimise black-box algorithms (Goldie et al., 2024; Lu et al., 2022), prompting LLMs to propose new functions (Lu et al., 2024), or using symbolic distillation of black-box functions to discover interpretable symbolic algorithms (Zheng et al., 2022). Due to the cost of meta-learning new algorithms, our study focuses on an intentionally diverse subset of meta-learning algorithms motivated by literature. Our study goes beyond any prior comparison of different meta-learning algorithms, improving our understanding of the field.
2. Based on our experimental results, we produce a set of recommendations for designing meta-learning pipelines. These proposals provide a 'snapshot' of the current state of the field, at the time of writing. These can help to make meta-learned algorithms as performant as possible within resource budgets.
   **Context:** Meta-learning experiments are very time-consuming and costly. For instance, Goldie et al. (2024) uses over 2 GPU-years of compute for meta-learning optimisers in small-scale RL environments, and Metz et al. (2022b) use over 4000 TPU-months to meta-learn a large versatile optimisation algorithm. By providing rough guidelines for where researchers should start their search, this work can help to reduce redundant initial experimentation while hopefully improving downstream performance. Our recommendations are based on experimental results, which are rooted in algorithms available from current literature. Different meta-learning algorithms are likely to improve and mature into the future, and thus our recommendations may need adaptation down the line.

# Seldonian Reinforcement Learning for Ad Hoc Teamwork

**Edoardo Zorzi, Alberto Castellini, Leonidas Bakopoulos, Georgios Chalkiadakis, Alessandro Farinelli**

**Keywords:** Offline Reinforcement Learning, Seldonian Algorithms, Ad Hoc Teamwork, Coordination, Trustworthy Reinforcement Learning

## Summary

Most offline RL algorithms return optimal policies but do not provide statistical guarantees on desirable behaviors. This could generate reliability issues in safety-critical applications, such as in some multiagent domains where agents, and possibly humans, need to interact to reach their goals without harming each other. In this work, we propose a novel offline RL approach, inspired by Seldonian optimization, which returns policies with good performance and statistically guaranteed properties with respect to predefined desirable behaviors. In particular, our focus is on Ad Hoc Teamwork settings, where agents must collaborate with new teammates without prior coordination. Our method requires only a pre-collected dataset, a set of candidate policies for our agent, and a specification about the possible policies followed by the other players—it does not require further interactions, training, or assumptions on the type and architecture of the policies. We test our algorithm in Ad Hoc Teamwork problems and show that it consistently finds reliable policies while improving sample efficiency with respect to standard ML baselines.

## Contribution(s)

1. We formalize the problem of offline Seldonian reinforcement learning in the context of Ad Hoc Teamwork, adding the assumption of having a candidate set of policies suitable for improvement and knowledge about the possible types of policies of the teammates. The goal is to select optimal policies with statistical guarantees with respect to desirable behaviors.
   **Context:** Seldonian policy optimization was first proposed in Thomas et al. (2019). An application to RL for diabetes management was introduced in the same work. The Ad Hoc Teamwork problem was introduced in Stone et al. (2010) with the goal of designing agents that can collaborate with new teammates without prior coordination. No other work formally defines the problem of offline Seldonian optimization for Ad Hoc Teamwork.
2. We provide an offline reinforcement learning algorithm based on Seldonian optimization for Ad Hoc Teamwork. The algorithm can solve multiagent scenarios due to an extended version of Doubly-Robust Importance Sampling (Jiang & Li, 2016), which estimates policy desirability by explicitly representing teammate policies and the transition model.
   **Context:** The Seldonian RL algorithm proposed in Thomas et al. (2019) cannot scale to large multiagent scenarios because of importance sampling inefficiency. We introduce, into the estimation process, knowledge about the Ad Hoc Teamwork setting to improve the algorithm's efficiency and allow it to scale. Other Seldonian algorithms in the literature (Satija et al., 2021; Thomas et al., 2015a) cannot deal with policy reliability in Ad Hoc Teamwork.
3. We evaluate our algorithm in three Ad Hoc Teamwork scenarios, showing that it is consistently reliable and efficient compared to standard ML baselines.
   **Context:** The evaluated domains are chain world (Chalkiadakis & Boutilier, 2003), extended blackjack (Sutton & Barto, 2018), and level-based foraging (Papoudakis et al., 2021).

# Offline Reinforcement Learning with Wasserstein Regularization via Optimal Transport Maps

**Motoki Omura, Yusuke Mukuta, Kazuki Ota, Takayuki Osa, Tatsuya Harada**

**Keywords:** Offline Reinforcement Learning, Deep Reinforcement Learning, Wasserstein Distance.

## Summary

Offline reinforcement learning (RL) aims to learn an optimal policy from a static dataset, making it particularly valuable in scenarios where data collection is costly, such as robotics. A major challenge in offline RL is distributional shift, where the learned policy deviates from the dataset distribution, potentially leading to unreliable out-of-distribution actions. To mitigate this issue, regularization techniques have been employed. While many existing methods utilize density ratio-based measures, such as the $f$-divergence, for regularization, we propose an approach that utilizes the Wasserstein distance, which is robust to out-of-distribution data and captures the similarity between actions. Our method employs input-convex neural networks (ICNNs) to model optimal transport maps, enabling the computation of the Wasserstein distance in a discriminator-free manner, thereby avoiding adversarial training and ensuring stable learning. Our approach demonstrates comparable or superior performance to widely used existing methods on the D4RL benchmark dataset. The code is available at https://github.com/motokiomura/Q-DOT.

## Contribution(s)

1. We introduce a novel regularization method with the Wasserstein distance via optimal transport maps for offline RL, eliminating the need for adversarial training and a discriminator through ICNNs.
   **Context:** Wu et al. (2019); Asadulaev et al. (2024) performed regularization using the Wasserstein distance in offline reinforcement learning through adversarial learning with a discriminator. Makkuva et al. (2020); Korotin et al. (2021b;a); Mokrov et al. (2021) modeled the Wasserstein distance in a discriminator-free manner using ICNNs in a non-RL domain.
2. We evaluate our proposed method on the D4RL benchmark dataset and find that it achieves performance comparable to or even surpassing that of widely used methods. Additionally, by comparing it with an adversarial training-based approach, we show that our discriminator-free method incorporates Wasserstein distance regularization more effectively for these tasks.
   **Context:** We compared our method with Kostrikov et al. (2022), which serves as a component of our approach, and a Wu et al. (2019)-based method that performs regularization using the Wasserstein distance via discriminator-based adversarial learning. By keeping the value function learning consistent across these existing methods and the proposed method, we fairly evaluated the effect of our proposed regularization on the policy.

# Bayesian Meta-Reinforcement Learning with Laplace Variational Recurrent Networks

**Joery A. de Vries, Jinke He, Mathijs M. de Weerdt, Matthijs T. J. Spaan**

**Keywords:** Variational Inference, Bayesian Reinforcement Learning, Meta-Reinforcement Learning, Uncertainty Estimation

## Summary

Meta-reinforcement learning trains a single reinforcement learning agent on a distribution of tasks to quickly generalize to new tasks outside of the training set at test time. From a Bayesian perspective, one can interpret this as performing amortized variational inference on the posterior distribution over training tasks. Among the various meta-reinforcement learning approaches, a common method is to represent this distribution with a point-estimate using a recurrent neural network. We show how one can augment this point estimate to give full distributions through the Laplace approximation, either at the start of, during, or after learning, without modifying the base model architecture. With our approximation, we are able to estimate distribution statistics (e.g., the entropy) of non-Bayesian agents and observe that point-estimate based methods produce overconfident estimators while not satisfying consistency. Furthermore, when comparing our approach to full-distribution based learning of the task posterior, our method performs similarly to variational baselines while having much fewer parameters.

## Contribution(s)

1. We formulate a probabilistic graphical model to match the practical design of memory-based meta-reinforcement learning agents, in order to perform uncertainty quantification through the Laplace approximation *without* retraining or architecture modifications.
   **Context:** Ours is an extension of the variational recurrent neural networks by Chung et al. (2015), *maximum a posteriori* policy optimization by Abdolmaleki et al. (2018), and adopts the control-as-inference framework (Levine, 2018).
2. We investigate how different assumptions on the posterior model over Markov decision processes interact with representation learning and uncertainty quantification of the recurrent neural network.
   **Context:** The agents trained with a recurrent neural network are non-Bayesian agents to which we try to apply a Bayesian approximation. Although we obtain a method for quantifying uncertainty in their learned representation, there is still a degree of misspecification.
3. When used as an alternative to the baseline variational recurrent network, we show that our method obtains similar performance.
   **Context:** This shows that our probabilistic formulation provides an alternative approximation for variational online learning while using fewer learnable parameters and without needing architecture modifications.
4. Our results show that the recurrent neural network representations learned by non-Bayesian meta-reinforcement learning agents, judging over multiple assumptions on the graphical model, produces overconfident estimators.
   **Context:** This extends prior insight by Xiong et al. (2021) that the representations of memory-based meta-reinforcement learning agents learn *inconsistent* estimators.

# Intrinsically Motivated Discovery of Temporally Abstract Graph-based Models of the World

**Akhil Bagaria, Anita de Mello Koch, Rafael Rodriguez-Sanchez, Sam Lobel, George Konidaris**

**Keywords:** Hierarchical RL, model-based RL, exploration.

## Summary

We seek to design reinforcement learning agents that build plannable models of the world that are abstract in both state and time. We propose a new algorithm to construct a *skill graph*; nodes in the skill graph represent abstract states and edges represent skill policies. Previous works that learn a skill graph use random sampling from the state-space and nearest-neighbor search—operations that are infeasible in environments with high-dimensional observations (for example, images). Furthermore, previous algorithms attempt to increase the probability of all edges (by repeatedly executing the corresponding skills) so that the resulting graph is robust and reliable everywhere. However, exhaustive coverage is infeasible in large environments, and agents should prioritize practicing skills that are more likely to result in higher reward. We propose a method to build skill graphs that aids exploration, without assuming state-sampling, distance metrics, or demanding exhaustive coverage.

## Contribution(s)

1. We provide an algorithm that learns a graph-based, plannable abstraction of the environment, even when the observations are high-dimensional; for example, images.
   **Context:** Prior work learned graph abstractions by assuming a distance metric over the state-space (for example, Bagaria et al. (2021b)), and hence cannot be easily applied to environments with image-based observations.
2. We use ideas from Intrinsic Motivation to design a graph construction algorithm that serves as a high-level exploration objective without needing to learn an accurate one-step model of the world—the drive to build the skill graph allows the agent to solve five challenging exploration problems, directly from pixels.
   **Context:** Most prior work in model-based exploration either operates in non-image based domains (for example, Sharma et al. (2020b)), or require the agent to learn one-step models (for example, Hafner et al. (2022); Mendonca et al. (2021)).
3. We provide a method for converting a goal-conditioned value function into a plannable abstract world model, which allows us to use dynamic programming to determine which option to execute at each state.
   **Context:** Lo et al. (2024) also study this problem, but they assume that option subgoals and initiation regions are provided to the agent; furthermore, they use the model for reward-shaping (Ng et al., 1999) rather than for option selection at decision time.

# An Optimisation Framework for Unsupervised Environment Design

**Nathan Monette, Alistair Letcher, Michael Beukman, Matthew T. Jackson, Alexander Rutherford, Alexander D. Goldie, Jakob N. Foerster**

**Keywords:** Optimisation, Environment Design, Reinforcement Learning, Robustness

## Summary

For reinforcement learning agents to be deployed in high-risk settings, they must achieve a high level of robustness to unfamiliar scenarios. One approach for improving robustness is unsupervised environment design (UED), a suite of methods that aim to maximise an agent's generalisability by training it on a wide variety of environment configurations. In this work, we study UED from an optimisation perspective, providing stronger theoretical guarantees for practical settings than prior work. Whereas previous methods relied on guarantees *if* they reach convergence, our framework employs a nonconvex-strongly-concave objective for which we provide a *provably convergent* algorithm in the zero-sum setting. We empirically verify the efficacy of our method, outperforming prior methods on two of three environments with varying difficulties.

## Contribution(s)

1. We provide a reformulation of UED that is strongly concave in the adversary's strategy, allowing for easier convergence.
   **Context:** Dennis et al. (2020)'s initial UED work PAIRED uses a nonconvex-nonconcave objective, which is known to be unstable in training (Wiatrak et al., 2020). Moreover, follow-up works such as (Chung et al., 2024) that improve PAIRED's level generator with generative models maintain this property.
2. We provide convergence guarantees for any score function that is a zero-sum game over the policy's negative return (e.g. regret or negative return).
   **Context:** Prior works in UED (Dennis et al., 2020; Jiang et al., 2021a) assert guarantees if the UED game reaches a saddle point, but fail to guarantee convergence to one. We propose a method that provably converges.
3. We provide an empirical evaluation of our methods on current UED benchmarks, using relevant optimisation heuristics and by introducing a new score function that generalises the work of Rutherford et al. (2024) to general deterministic RL environments.
   **Context:** Learnability (Rutherford et al., 2024) requires a binary-outcome environment.

# Epistemically-guided forward-backward exploration

**Núria Armengol Urpí, Marin Vlastelica, Georg Martius, Stelian Coros**

**Keywords:** unsupervised RL, exploration, zero-shot, epistemic uncertainty, ensemble

## Summary

The goal of zero-shot RL is to provide algorithms for recovering optimal policies for all possible reward functions given interaction data with the environment. Naturally, how well we can recover the optimal policies highly depends on the quality of the data used to learn them. Up until now, most algorithms leverage decoupled exploration policies for collecting data in order to learn a generalist representation of all optimal policies. A central argument to this paper is that the exploration policy should not be completely decoupled from the zero-shot algorithm and should try to minimize the uncertainty that the algorithm has of its representations. We frame the exploration problem for zero-shot RL as minimization of the epistemic uncertainty on the learned value functions, and realize this in the case of well familiar algorithm, forward-backward (FB) representations. Crucially, in several empirical settings, using an exploration policy that maximizes the cumulative epistemic uncertainty of the FB representation leads to significant improvements of the algorithm's sample complexity. This enables us to learn well-performing policies fast, with fewer amount of data than other exploration approaches.

## Contribution(s)

1. This paper phrases the exploration problem for zero-shot RL as uncertainty minimization of a posterior over occupancy measures for a particular representation of an occupancy measure. The main difference to previous work is that, while previous work considers completely off-policy exploration algorithms to collect data, this paper considers the uncertainty of the model for data collection in an unsupervised RL setting.
   **Context:** The representation for occupancy measure used is the FB-representation (Touati & Ollivier, 2021) which encodes all optimal policies. We use an ensemble method approximation to the posterior distribution. Crucially, because of non-uniqueness, the FB representation does not allow simple modeling of the posterior uncertainty over FB via ensemble disagreement – there is a necessity of having a single $B$ representation in order to have an informative notion of uncertainty. Furthermore, the $F$-uncertainty is projected to the more practical uncertainty over $Q$-functions for particular latent policy conditioning $z$.
2. We introduce an efficient algorithm for exploration tailored to forward-backward (FB) representations which can be seen as a variant of *uncertainty sampling* (Lewis & Gale, 1994).
   **Context:** The algorithm relies on sampling a posterior-mean greedy policy $\pi_z$ which has highest uncertainty in the predictive posterior distribution for a particular state $s$ and executing it in the environment. This exploration strategy, while simple and not considering correlation in uncertainty reduction across all policies $\pi_z, z \in \mathcal{Z}$, is a surprisingly efficient method for exploration in FB representations.
3. Experimental validation of proposed exploration on several continuous control environments from the DeepMind Control suite (Tassa et al., 2018) in the online learning setting, where we evaluate zero-shot performance on different reward functions within several environments.
   **Context:** There is no notion of exploitation in the unsupervised RL setting, therefore there is no need to balance the exploration-exploitation trade-off when collecting data. This setup is fundamentally different than single-task online learning, where typically we balance an intrinsic exploration signal or noise with the extrinsic task reward.

# Rethinking the Foundations for Continual Reinforcement Learning

**Esraa Elelimy, David Szepesvari, Martha White , Michael Bowling**

**Keywords:** Continual reinforcement learning, hindsight rationality, history process, MDPs

## Summary

In the traditional view of reinforcement learning, the agent's goal is to find an optimal policy that maximizes its expected sum of rewards. Once the agent finds this policy, the learning ends. This view contrasts with *continual reinforcement learning*, where learning does not end, and agents are expected to continually learn and adapt indefinitely. Despite the clear distinction between these two paradigms of learning, much of the progress in continual reinforcement learning has been shaped by foundations rooted in the traditional view of reinforcement learning. In this paper, we first examine whether the foundations of traditional reinforcement learning are suitable for the continual reinforcement learning paradigm. We identify four key pillars of the traditional reinforcement learning foundations that are antithetical to the goals of continual learning: the Markov decision process formalism, the focus on atemporal artifacts, the expected sum of rewards as an evaluation metric, and episodic benchmark environments that embrace the other three foundations. We then propose a new formalism that sheds the first and the third foundations and replaces them with the history process as a mathematical formalism and a new definition of deviation regret, adapted for continual learning, as an evaluation metric. Finally, we discuss possible approaches to shed the other two foundations.

## Contribution(s)

1. We identify four foundational principles and practices that shape and constrain our thinking about RL. We argue that these foundations, shaped by the traditional framing of RL, are antithetical to the purported goals of continual reinforcement learning and may be holding us back from making progress toward continual learning.
   **Context:** Most of our arguments are in alignment with the constraints that arise under the big world hypothesis (Javed & Sutton, 2024). Previous work by Abel et al. (2024b) has discussed three dogmas that shape most reinforcement learning research. The second dogma overlaps with the second foundation that we argue against as part of the foundations of traditional reinforcement learning. In this work, we identify three additional problematic foundations in the traditional RL framing and, for completeness, we reiterate some of the arguments against this second dogma as well.
2. We present a new formalism that replaces two of the foundations with the history process as a mathematical formalism and deviation regret as an evaluation metric.
   **Context:** The history process foundation is built on earlier work by Bowling et al. (2023) and Hutter (2000), and the deviation regret is an extension to earlier work by Morrill et al. (2021b). The earlier work on deviation regret by Morrill et al. (2021b) focused on settings where the history process can be repeated, such as in extensive-form games. In this work, we extend the notion of deviation regret to the continual learning setting and provide some theoretical analysis on deviation regret estimation in the continual learning setting.
3. We present experimental results suggesting that the current RL algorithms fail to learn continually and that our proposed measure of evaluation can evaluate those failures.
   **Context:** Platanios et al. (2023) showed similar results for agents failing to learn continually, which aligns with our experimental findings. We extend those results to show the utility of deviation regret as an evaluation measure when agents fail to learn.

# Modelling human exploration with light-weight meta reinforcement learning algorithms

**Thomas D. Ferguson, Alona Fyshe, Adam White**

**Keywords:** exploration, multi-arm bandit, IDBD, meta-learning, non-stationary

## Summary

Learning in non-stationary environments can be difficult. Although many algorithmic approaches have been developed, often methods struggle with different forms of non-stationarity such as gradually changing versus suddenly changing contexts. Luckily, humans can learn effectively under a variety of conditions and using human learning could be revealing. In the present work, we investigated if a stateless variant of the IDBD algorithm (Mahmood et al., 2012; Sutton, 1992), which has previously shown success in bandit-like tasks (Linke et al., 2020), can model human exploration. We compared stateless IDBD to two algorithms that are frequently used to model human exploration (a standard Q-learning algorithm and a Kalman filter algorithm). We examined the ability of these three algorithms to fit human choices and to replicate human learning within three different bandits: (1) non-stationary volatile which changed suddenly, (2) non-stationary drifting which changed gradually, and (3) stationary. In these three bandits, we found that stateless IDBD provided the best fit of the human data and was best able to replicate different aspects of human learning. We also found that when fit to the human data, differences in the hyperparameters of stateless IDBD across the three bandits may explain how humans learn effectively across contexts. Our results demonstrate that stateless IDBD can account for different types of non-stationarity and model human exploration effectively. Our findings highlight that taking inspiration from algorithms used with artificial agents may provide further insights into both human learning and inspire the development of algorithms for use in artificial agents.

## Contribution(s)

1. Our work is the first to investigate a light-weight, meta-learning algorithm from reinforcement learning (IDBD) as a potential computational model of human exploration. Recovery of IDBD parameters and simulation results from our human data provides suggestive evidence that people modulate their learning rates in a similar manner to IDBD.
   **Context:** Our work may be limited to the bandit setting, as human data in multi-stage decision making tasks is typically modelled using hybrid model-free/model-based (e.g., successor representation) approaches (Momennejad et al., 2017).
2. Although prior work has shown IDBD-based agents can automatically and continually adapt step-sizes to improve performance in simulation, we are the first to show IDBD can do the same with human exploration data (i.e., a sequence of actions and rewards generated by people performing bandit tasks).
   **Context:** IDBD-inspired agents have been used in supervised learning tasks (Sutton, 1992; Mahmood et al., 2012), bandit tasks (Linke et al., 2020), MDPs (Mcleod et al., 2021; Kearney et al., 2018; Javed et al., 2024; Jacobsen et al., 2019), and even to help predict data from real robots (Mahmood et al., 2012; Kearney et al., 2018; Jacobsen et al., 2019)
3. Our analysis indicates that IDBD matches human data better when compared to a Q-learning and a Kalman filter algorithm which were used as baselines (Daw et al., 2006; Hassall et al., 2019).
   **Context:** Our results are limited to three tasks and a moderate number of human participants. It is always possible that different tasks or a larger number of participants could produce different conclusions. We did not exhaustively study all computational models proposed in the literature, but instead focused on two: a Q-learning algorithm (Hassall et al., 2019) and a Kalman filter algorithm (Daw et al., 2006)

# Zero-Shot Reinforcement Learning Under Partial Observability

**Scott Jeen, Tom Bewley, Jonathan M. Cullen**

**Keywords:** Zero-shot RL, Behaviour Foundation Models, POMDPs.

## Summary

Recent work has shown that, under certain assumptions, zero-shot reinforcement learning (RL) methods can generalise to *any* unseen task in an environment after reward-free pre-training. Access to Markov states is one such assumption, yet, in many practical applications, the Markov state is only *partially observed* via unreliable or incomplete observations. Here, we explore how the performance of standard zero-shot RL methods degrades when subjected to partially observability, and show that, as in single-task RL, memory-based architectures are an effective remedy. We evaluate our *memory-based* zero-shot RL methods in domains where we simulate unreliable states by adding noise or dropping them randomly, and in domains where we simulate incomplete observations by changing the dynamics between training and testing rewards without communicating the change to the agent. In these settings, our proposals show improved performance over memory-free baselines, which we pay for with slower, less stable training dynamics.

## Contribution(s)

1. We explore the empirical failure modes of state-of-the-art zero-shot RL methods (specifically forward-backward representations, or FB) given partially observed (noisy) states.
   **Context:** None
2. We present a new architecture called FB with memory (FB-M) which has a memory-based forward model $F$, backward model $B$ and policy $\pi$. Though we develop our method within the FB framework, our proposals are fully compatible with other zero-shot RL methods.
   **Context:** Prior zero-shot RL methods, including FB (Touati & Ollivier, 2021) and USF-based HILP (Borsa et al., 2018; Park et al., 2024b), are memory-free.
3. We show that, in aggregate, FB-M outperforms memory-free FB and HILP, as well as a naïve observation-stacking baseline, in domains where the states are noisy or randomly dropped, or where there is a change in dynamics function between training and testing.
   **Context:** None
4. We report better performance when the memory model is a GRU than when it is a transformer or S4d model.
   **Context:** This aligns with Morad et al. (2023)'s finding that GRUs were the most performant memory model on POPGym, a partially observed single-task RL benchmark.

# Building Sequential Resource Allocation Mechanisms without Payments

**Sihan Zeng, Sujay Bhatt, Alec Koppel, Sumitra Ganesh**

**Keywords:** Sequential resource allocation, mechanism design, incentive compatibility.

## Summary

We study allocating limited divisible resources to agents who submit requests for the resources one or multiple times over a finite horizon. This is referred to as the sequential resource allocation problem, as irrevocable allocations need to be made as the requests arrive, without observations on the future requests. Existing works on sequential resource allocation (in the payment-free setting) mainly focus on optimizing social welfare and design mechanisms under the assumption that the agents make truthful requests. Such mechanisms can be easily exploitable – strategic agents may misreport their requests and inflate their allocations. Our aim in this work is to design sequential resource allocation mechanisms that balance the competing objectives of social welfare maximization (promoting the overall agent satisfaction) and incentive compatibility (ensuring that the agents do not have incentives to misreport). We do not design these mechanisms from scratch. As the incentive compatible mechanism design problem has been well studied in the *one-shot* setting (horizon length equals one), we propose a general *meta-algorithm* of transforming a one-shot mechanism into its sequential counterpart. The meta-algorithm can plug in any one-shot mechanism and approximately carry over the properties that the one-shot mechanism already satisfies to the sequential setting. We establish theoretical results validating these claims and also illustrate the superior performance of the proposed method through numerical simulations.

## Contribution(s)

1. We propose a meta-algorithm, which we name **S**equential **A**llocation **M**eta **A**lgorithm (`SAMA`), which can be regarded as a general framework for reducing a sequential resource allocation problem into a series of one-shot problems. The key feature of `SAMA` is that it accounts for past allocation and unobserved future requests – agents with greater past allocations are more discounted against in the current round, and resources are withheld for future requests based on a confidence bound. We mathematically show that if the one-shot mechanism optimizes NSW and/or achieves incentive compatibility (IC) in the one-shot sense, `SAMA` approximately carries over the properties to the sequential setting. To our knowledge, this is the first time such a result has been established for a sequential mechanism in the payment-free setting.
   **Context:** Prior papers on sequential resource allocation do not consider achieving IC and assume that the agents report their requests truthfully. The existing work that considers optimizing IC jointly with other metrics including social welfare and efficiency is only for the one-shot setting, in which the supplier fully observes all requests before making an allocation.
2. We numerically illustrate the superior performance of `SAMA` and its approximate NSW and IC preserving properties, with a few established one-shot mechanisms as the building block. Specifically, we plug-in 1) the Proportional Fairness (PF) mechanism, which achieves the maximum possible NSW but severely violates IC, 2) the Partial Allocation (PA) mechanism, designed by Cole et al. (2013) to be exactly IC at the cost of a substantial reduction to NSW, 3) ExS-Net, which is a learned neural-network-parameterized mechanism proposed in Zeng et al. (2024b) that achieves near-optimal NSW and approximate IC simultaneously.
   **Context:** None.

# From Explainability to Interpretability: Interpretable Reinforcement Learning Via Model Explanations

**Peilang Li, Umer Siddique, Yongcan Cao**

**Keywords:** Deep Reinforcement Learning, Interpretable Reinforcement Learning, Explainable Reinforcement Learning, Shapley Values.

## Summary

Deep reinforcement learning (RL) has shown remarkable success in complex domains, however, the inherent black box nature of deep neural network policies raises significant challenges in understanding and trusting the decision-making processes. While existing explainable RL methods provide local insights, they fail to deliver a global understanding of the model, particularly in high-stakes applications. To overcome this limitation, we propose a novel model-agnostic framework that bridges the gap between explainability and interpretability by leveraging Shapley values to transform complex deep RL policies into transparent representations. The proposed approach, SILVER (Shapley value-based Interpretable poLicy Via Explanation Regression) offers two key contributions: a novel approach employing Shapley values to policy interpretation beyond local explanations, and a general framework applicable to off-policy and on-policy algorithms. We evaluate SILVER with three existing deep RL algorithms and validate its performance in three classic control environments. The results demonstrate that SILVER not only preserves the original models' performance but also generates more stable interpretable policies.

## Contribution(s)

1. This paper presents a novel framework to derive interpretable policies from explainable methods.
   **Context:** Prior work focused on generating explanations in Reinforcement Learning without deriving an interpretable policy from it. (Beechey et al., 2023)
2. This framework generates highly transparent, interpretable policies while maintaining model performance.
   **Context:** It overturns the conventional assumption that there must be a trade-off between interpretability and performance.
3. This model-agnostic framework is applicable to both off-policy and on-policy reinforcement learning algorithms.
   **Context:** Prior works are mostly model-specific, limiting their ability to generalize across diverse RL scenarios.

# Joint-Local Grounded Action Transformation for Sim-to-Real Transfer in Multi-Agent Traffic Control

**Justin Turnau, Longchao Da, Khoa Vo, Ferdous Al Rafi, Shreyas Bachiraju, Tiejin Chen, Hua Wei**

**Keywords:** Traffic Signal Control, Multi-Agent Reinforcement Learning, Sim-to-Real Transfer

## Summary

Traffic Signal Control (TSC) is essential for managing urban traffic flow and reducing congestion. Reinforcement Learning (RL) offers an adaptive method for TSC by responding to dynamic traffic patterns, with multi-agent RL (MARL) gaining traction as intersections naturally function as coordinated agents. However, due to shifts in environmental dynamics, implementing MARL-based TSC policies in the real world often leads to a significant performance drop, known as the sim-to-real gap. Grounded Action Transformation (GAT) has successfully mitigated this gap in single-agent RL for TSC, but real-world traffic networks, which involve numerous interacting intersections, are better suited to a MARL framework. In this work, we introduce JL-GAT, an application of GAT to MARL-based TSC that balances scalability with enhanced grounding capability by incorporating information from neighboring agents. JL-GAT adopts a decentralized approach to GAT, allowing for the scalability often required in real-world traffic networks while still capturing key interactions between agents. Comprehensive experiments on various road networks and ablation studies demonstrate the effectiveness of JL-GAT.

## Contribution(s)

1. We introduce Joint-Local Grounded Action Transformation (JL-GAT), a scalable framework for bridging the sim-to-real gap in MARL-based traffic signal control that incorporates state and action information from neighboring agents into Grounded Action Transformation (GAT) models using a sensing radius.
   **Context:** None
2. To the best of our knowledge, we are the first to apply Grounded Action Transformation (GAT) to the multi-agent setting, introducing two natural applications of GAT alongside our proposed method, JL-GAT.
   **Context:** None
3. We introduce the cascading invalidation effect, a novel challenge in JL-GAT that arises when integrating state and action information from nearby agents, and propose both a direct solution and an alternative approach that effectively mitigates the issue.
   **Context:** None
4. We conduct thorough empirical evaluations of JL-GAT in the domain of multi-agent traffic signal control, demonstrating its effectiveness in reducing the sim-to-real gap.
   **Context:** None

# Foundation Model Self-Play: Open-Ended Strategy Innovation via Foundation Models

**Aaron Dharna, Cong Lu, Jeff Clune**

**Keywords:** open-ended learning, self-play, quality-diversity, foundation models, policy search

## Summary

Self-play (SP) algorithms try to harness multi-agent dynamics by pitting agents against ever-improving opponents to learn high-quality solutions. However, SP often fails to learn diverse solutions and can get stuck in locally optimal behaviors. We introduce Foundation-Model Self-Play (FMSP), a new direction that leverages the code-generation capabilities and vast knowledge of foundation models (FMs) to overcome these challenges. We propose a *family* of approaches: (1) **Vanilla Foundation-Model Self-Play (vFMSP)** continually refines agent policies via competitive self-play; (2) **Novelty-Search Self-Play (NSSP)** builds a diverse population of strategies, ignoring performance; and (3) the most promising variant, **Quality-Diversity Self-Play (QDSP)**, creates a diverse set of high-quality policies by combining elements of NSSP and vFMSP. We evaluate FMSPs in Car Tag, a continuous-control pursuer-evader setting, and in Gandalf, a simple AI safety simulation in which an attacker tries to jailbreak an LLM's defenses. In Car Tag, FMSPs explore a wide variety of reinforcement learning, tree search, and heuristic-based methods, to name just a few. In terms of discovered policy quality, QDSP and vFMSP surpass strong human-designed strategies. In Gandalf, FMSPs can successfully automatically red-team an LLM, breaking through and jailbreaking six different, progressively stronger levels of defense. Furthermore, FMSPs can automatically proceed to patch the discovered vulnerabilities. Overall, FMSP and its many possible variants represent a promising new research frontier of improving self-play with foundation models, opening fresh paths toward more creative and open-ended strategy discovery.

## Contribution(s)

1. We propose *foundation-model self-play* (FMSP), a new family of policy search algorithms that combine the implicit curriculum of multi-agent self-play (Silver et al., 2016; Tesauro, 1994) with foundation-model code generation (Bommasani et al., 2021; Liang et al., 2022) to create high-quality policies. We introduce three FMSP variants each inspired by traditional search—Vanilla FMSP, Novelty Search Self Play, and Quality-Diversity Self Play. Each FMSP variant explores different exploration and exploitation tradeoffs.
   **Context:** Prior work has shown that foundation models can generate single-agent code-based policies (Liang et al., 2022; Wang et al., 2023a), but this is the first work to co-evolve code-based agents in multi-agent settings with FMs powering the search. vFMSP is a pure exploitation-driven algorithm, analogous to FM-driven single-objective optimization (Wang et al., 2023a), but here in multi-agent self-play; NSSP is a pure exploration-driven algorithm that leverages FM's models of human interestingness (Zhang et al., 2023) to generate a diverse set of policies; and QDSP is a hybrid approach combining vFMSP's hill-climbing with NSSP's novelty-seeking to create the first MAP-Elites algorithm (Mouret & Clune, 2015) where the practitioner need not define the dimensions of interest.
2. FMSPs discover diverse and effective strategies in two multi-agent tasks: (i) Car Tag (Isaacs & Corporation, 1951), a continuous-control pursuer-evader task, and (ii) Gandalf, a novel AI-safety puzzle where an attacker jailbreaks an LLM and the defender patches exploits.
   **Context:** We thus show the benefits of FM-powered SP on diverse domains, and highlight their benefit for traditional control tasks as well as AI safety.

# Action Mapping for Reinforcement Learning in Continuous Environments with Constraints

**Mirco Theile, Lukas Dirnberger, Raphael Trumpp, Marco Caccamo, Alberto L. Sangiovanni-Vincentelli**

**Keywords:** Action masking, constrained MDPs, continuous action space, deep reinforcement learning

## Summary

Deep reinforcement learning (DRL) has had success across various domains, but applying it to environments with constraints remains challenging due to poor sample efficiency and slow convergence. Recent literature explored incorporating model knowledge to mitigate these problems, particularly using models that assess the feasibility of proposed actions. However, integrating feasibility models efficiently into DRL pipelines in environments with continuous action spaces is non-trivial. We propose a novel DRL training strategy utilizing *action mapping* that leverages feasibility models to streamline the learning process. By decoupling the learning of feasible actions from policy optimization, action mapping allows DRL agents to focus on selecting the optimal action from a reduced feasible action set. We demonstrate that action mapping significantly improves training performance in two constrained environments with continuous action spaces, especially with imperfect feasibility models.

## Contribution(s)

1. In this paper, we develop and implement the action mapping (AM) framework for DRL to efficiently incorporate feasibility models during training. In AM, the training is split into two steps. First, a feasibility policy is trained to generate all feasible actions given a state by leveraging the feasibility model. Second, an objective policy learns to select the optimal action among these pretrained feasible actions.
   **Context:** The AM framework was originally conceptualized by Theile et al. (2024). However, they only focussed on the feasibility policy, omitting the objective policy and thus leaving its practical benefits in DRL unexplored. In this paper, we refine their feasibility policy training and formulate the training procedure for the objective policy.
2. Using perfect and approximate feasibility models with AM-PPO and AM-SAC implementations, we demonstrate AM's effectiveness in constrained environments. Empirical comparison with Lagrangian methods and action replacement, resampling, and projection highlights superior performance, especially with approximate models.
   **Context:** While action replacement, resampling, and projection utilize the feasibility models, the Lagrangian methods are model-free and thus have an innate disadvantage. However, they were added to also compare with model-free approaches.
3. Additionally, we showcase AM's ability to express multi-modal action distributions, enhancing exploration and learning performance.
   **Context:** Commonly, the output of a DRL policy is parameterizing a single-mode Gaussian, which can be disadvantageous when there are disconnected sets of feasible actions. AM allows the agent to effectively produce multi-mode Gaussians in the action space, allowing it to explore actions in disconnected sets of feasible actions.

# Sampling from Energy-based Policies using Diffusion

**Vineet Jain, Tara Akhound-Sadegh, Siamak Ravanbakhsh**

**Keywords:** Energy-based policies, Boltzmann policies, diffusion models.

## Summary

Energy-based policies offer a flexible framework for modeling complex, multimodal behaviors in reinforcement learning (RL). In maximum entropy RL, the optimal policy is a Boltzmann distribution derived from the soft Q-function, but direct sampling from this distribution in continuous action spaces is computationally intractable. As a result, existing methods typically use simpler parametric distributions, like Gaussians, for policy representation — limiting their ability to capture the full complexity of multimodal action distributions. In this paper, we introduce a diffusion-based approach for sampling from energy-based policies, where the negative Q-function defines the energy function. Based on this approach, we propose an actor-critic method called Diffusion Q-Sampling (DQS) that enables more expressive policy representations, allowing stable learning in diverse environments. We show that our approach enhances sample efficiency in continuous control tasks and captures multimodal behaviors, addressing key limitations of existing methods.

## Contribution(s)

1. We develop a novel actor-critic reinforcement learning algorithm such that the policy samples actions from the Boltzmann distribution of the Q-function. We achieve this by using a diffusion model to parameterize the policy that explicitly learns the score function of the target Boltzmann density.
   **Context:** Boltzmann policies are a popular choice in discrete action spaces. However, sampling from these policies in continuous action spaces is generally intractable. Prior work (Psenka et al., 2023) used Langevin sampling to address this challenge. Other applications of diffusion models (Wang et al., 2024) backpropagate the gradient through the entire diffusion chain to maximize Q-values. To the best of our knowledge, our method is the first to use diffusion models to explicitly sample from Boltzmann policies.
2. Experiments on continuous control tasks demonstrate improved sample efficiency of our method compared to relevant baselines.
   **Context:** We observe higher returns with fewer number of environment interactions (compared to our baselines) on a majority of tasks.
3. We demonstrate that our proposed method can learn multimodal behaviors in maze navigation tasks.
   **Context:** Our setup consists of a maze with two possible goals. Multimodality in this context refers to the ability of an agent to reach both goals from some initial state, and discover multiple paths (if they exist) to a goal. We qualitatively examine the trajectories of a trained agent and compare them with respect to goal coverage and diversity of paths.

# Multiple-Frequencies Population-Based Training

**Waël Doulazmi, Auguste Lehuger, Marin Toromanoff, Valentin Charraut, Thibault Buhet, Fabien Moutarde**

**Keywords:** Hyperparameter Optimization, Greediness, Reinforcement Learning

## Summary

Reinforcement Learning's high sensitivity to hyperparameters is a source of instability and inefficiency, creating significant challenges for practitioners. Hyperparameter Optimization (HPO) algorithms have been developed to address this issue, among them Population-Based Training (PBT) stands out for its ability to generate hyperparameters schedules instead of fixed configurations. PBT trains a population of agents, each with its own hyperparameters, frequently ranking them and replacing the worst performers with mutations of the best agents. These intermediate selection steps can cause PBT to focus on short-term improvements, leading it to get stuck in local optima and eventually fall behind vanilla Random Search over longer timescales. This paper studies how this greediness issue is connected to the choice of *evolution frequency*, the rate at which the selection is done. We propose Multiple-Frequencies Population-Based Training (MF-PBT), a novel HPO algorithm that addresses greediness by employing sub-populations, each evolving at distinct frequencies. MF-PBT introduces a migration process to transfer information between sub-populations, with an asymmetric design to balance short and long-term optimization.

## Contribution(s)

1. We investigate the impact of evolution frequency on PBT and its connection to greediness.
   **Context:** PBT (Jaderberg et al., 2017) introduces a parameter, denoted $t_{\text{ready}}$, which controls the evolution frequency of its genetic process. Previous extensions of PBT (Parker-Holder et al., 2020; 2021; Franke et al., 2021; Dalibard & Jaderberg, 2021; Wan et al., 2022) employ a $t_{\text{ready}}$ parameter, but don't study or comment its impact on performance. We show it can be used to control PBT's optimization horizon, avoiding greedy behaviors that makes PBT weak for long-term performance. The greediness of PBT was identified in the original PBT paper (Jaderberg et al., 2017), and in FIRE PBT (Dalibard & Jaderberg, 2021). But we propose a novel scope to analyse it: evolution frequency.
2. We introduce MF-PBT to mitigate the greediness issue.
   **Context:** FIRE PBT (Dalibard & Jaderberg, 2021) is the only other attempt at solving the greediness issue. We propose a simpler approach, that is very close to the original PBT algorithm. We make our work reproducible by publishing the code and addressing all the implementation details in the paper.
3. We evaluate MF-PBT and ablate its components. We build an experimental setup that enables us to exhibit the greediness issues of population-based approaches, and show that MF-PBT effectively mitigates greediness.
   **Context:** Our experiments rely on the Brax (Freeman et al., 2021) framework, whose speed enables to perform experiments on the billion steps scale. MF-PBT does not claim to be a SOTA approach to HPO. Our contribution is to isolate, explain and mitigate an important weakness of PBT, which is a popular HPO method for RL.
4. We empirically show how population-based methods can leverage stochasticity in RL training to significantly improve performance, even without tuning hyperparameters.
   **Context:** Performance gains are usually associated to the effective optimization of hyperparameters. We show that beyond HPO, PBT can be used in a *variance-exploitation* mode, to bring significant performance gains on an already-tuned hyperparameter configuration. We further show that PBT still exhibits greediness in this mode and that MF-PBT is a better solution.

# TransAM: Transformer-Based Agent Modeling for Multi-Agent Systems via Local Trajectory Encoding

**Conor Wallace, Umer Siddique, Yongcan Cao**

**Keywords:** Multi-Agent Systems, Agent Modeling, Transformer Networks, Policy Representation, Adaptive Learning.

## Summary

Agent modeling is a critical component in developing effective policies within multi-agent systems, as it enables agents to form beliefs about the behaviors, intentions, and competencies of others. Many existing approaches assume access to other agents' episodic trajectories, a condition often unrealistic in real-world applications. Consequently, a practical agent modeling approach must learn a robust representation of the policies of the other agents based only on the local trajectory of the controlled agent. In this paper, we propose `TransAM`, a novel transformer-based agent modeling approach to encode local trajectories into an embedding space that effectively captures the policies of other agents. We evaluate the performance of the proposed method in cooperative, competitive, and mixed multi-agent environments. Extensive experimental results demonstrate that our approach generates strong policy representations, improves agent modeling, and leads to higher episodic returns.

## Contribution(s)

1. We eliminate the need for access to other agents' trajectories at inference time by learning a latent policy representation derived solely from the local trajectory of the controlled agent.
   **Context:** It is common for agent modeling methods to assume access to other agent information at execution time (He & Boyd-Graber, 2016; Grover et al., 2018; Jing et al., 2024).
2. By treating the local trajectory of the controlled agent as a temporal sequence, we use a transformer to model long-range dependencies and identify key moments that characterize interactions with other agents. This is in contrast to previous methods that rely on MLPs or RNNs without attention over extended time horizons.
   **Context:** Other methods typically construct either an MLP-based agent model (He & Boyd-Graber, 2016), or a recurrent agent model (Papoudakis et al., 2021), which do not take into account the full context of the agent's trajectory throughout the episode.
3. To address the data demands of transformers, we train the agent model and the controlled agent's policy jointly in an online setting, ensuring access to a diverse dataset for enhanced performance.
   **Context:** Other promising transformer-based agent modeling approaches, such as (Jing et al., 2024) are based on an offline reinforcement learning setting wherein a pretraining phase is used to learn an initial prior for the task. In contrast, we aim to train the agent model and the policy jointly from scratch.

# Chargax: A JAX Accelerated EV Charging Simulator

**Koen Ponse , Jan Felix Kleuker, Aske Plaat, Thomas Moerland**

**Keywords:** Jax, EV Charging, Gym Environment, Reinforcement Learning, Benchmarking

## Summary

Deep Reinforcement Learning can play a key role in addressing sustainable energy challenges. For instance, many grid systems are heavily congested, highlighting the urgent need to enhance operational efficiency. However, reinforcement learning approaches have traditionally been slow due to the high sample complexity and expensive simulation requirements. While recent works have effectively used GPUs to accelerate data generation by converting environments to JAX, these works have largely focussed on classical toy problems. This paper introduces Chargax, a JAX-based environment for realistic simulation of electric vehicle charging stations designed for accelerated training of RL agents. We validate our environment in a variety of scenarios based on real data, comparing reinforcement learning agents against baselines. Chargax delivers substantial computational performance improvements of over 100x-1000x over existing environments. Additionally, Chargax' modular architecture enables the representation of diverse real-world charging station configurations.

## Contribution(s)

(i) This paper presents Chargax, an open-source EV charging environment written in JAX
**Context:** Chargax could be used as a high-performance test bed for reinforcement learning benchmarking, or to develop better control algorithms for EV charging.

(ii) Comparisons in performance are made with previously existing EV simulators for RL that demonstrate Chargax decreases training times by a factor of 100x or more.
**Context:** Prior work established EV charging simulators for RL that did no leverage the GPU

(iii) We perform additional experiments validating reinforcement learning training in a variety of scenarios, data distributions shifts, and reward objectives.
**Context:** None

(iv) We create an explicit split in the state space which highlights the interchangeable parts in the Chargax environment. This modularity allows representation of diverse real-world charging station configurations and scenarios.
**Context:** Prior work often used this split implicitly, and allow for less customisability

# Towards Improving Reward Design in RL: A Reward Alignment Metric for RL Practitioners

**Calarina Muslimani, Kerrick Johnstonbaugh, Suyog Chandramouli, Serena Booth, W. Bradley Knox, Matthew E. Taylor**

**Keywords:** Reinforcement Learning, Reward Design, Alignment, Human-AI Interaction, Human-in-the-loop

## Summary

Reinforcement learning (RL) agents are fundamentally limited by the quality of the reward functions they learn from, yet reward design is often overlooked under the assumption that a well-de ned reward is readily available. However, this is rarely the case in real-world applications, and reward design is a challenging endeavor: sparse rewards can hinder learning, while dense rewards may increase the risk of misspeci cation. Reward evaluation is equally problematic: *how do we know if a reward function is correctly specified?* In our work, we address this challenge by focusing on *reward alignment* assessing whether a reward function accurately encodes the preferences of a human stakeholder. As a concrete measure of reward alignment, we introduce the Trajectory Alignment Coef cient. This metric quanti es the similarity between a human stakeholder's ranking of trajectory distributions and those induced by a given reward function. We validate the usefulness of this metric theoretically and through a user study.

## Contribution(s)

1. This paper introduces the Trajectory Alignment Coef cient, a metric to evaluate reward alignment the extent to which a reward function encodes the preferences of a human stakeholder. This metric only requires access to human preferences over trajectory distributions.
   **Context:** Prior work (Gleave et al., 2021; Wulfe et al., 2022) has proposed reward distance metrics, which can be treated as reward alignment metrics if one of the reward functions accurately re ects the preferences of a human stakeholder. Other work (Brown et al., 2021) has proposed alignment veri cation of reward functions. However, these metrics assume access to ground-truth reward or value functions which our metric does not require.
2. Through an 11 person user study, we demonstrate that the Trajectory Alignment Coef cient can assist RL practitioners (self-identi ed) with reward selection, compared to relying solely on inspection of the reward function de nition. Speci cally, we found the following statistically signi cant results: (1) access to the Trajectory Alignment Coef cient during reward selection reduced perceived cognitive workload by 1.5x, (2) was preferred by 82% of participants over the Reward Only condition, and (3) increased the success rate of selecting reward functions that produced performant policies by 41% (compared to unselected alternative rewards).
   **Context:** The user study results presented in this work are in the context of the Hungry-Thirsty domain (Singh et al., 2009). This test-bed has been shown to be particularly challenging for reward design (Booth et al., 2023). Further, we speci cally do not include Gleave et al. (2021); Wulfe et al. (2022); Brown et al. (2021) as conditions in the user study because they are not directly comparable due to the difference in necessary assumptions.
3. We prove that the Trajectory Alignment Coef cient is invariant to potential-based reward shaping and positive linear transformations of the reward function if and only if the metric considers trajectory distributions that share the same start state distribution.
   **Context:** Invariance to these reward transformations is a common property of reward evaluation metrics proposed in previous works (Gleave et al., 2021; Wulfe et al., 2022).

# Optimistic critics can empower small actors

**Olya Mastikhina, Dhruv Sreenivas, Pablo Samuel Castro**

**Keywords:** Deep reinforcement learning, actor-critic, asymmetric actor-critics, exploration, value underestimation, data collection

## Summary

Actor-critic methods have been central to many of the recent advances in deep reinforcement learning. The most common approach is to use *symmetric* architectures, whereby both actor and critic have the same network topology and number of parameters. However, recent works have argued for the advantages of *asymmetric* setups, specifically with the use of smaller actors. We perform broad empirical investigations and analyses to better understand the implications of this.

## Contribution(s)

1. We show that reducing the size of the actor in actor-critic methods can lead to degraded performance and increased overfitting in the critic.
   **Context:** Prior work suggests that actors require less capacity than critics in actor-critic algorithms (Mysore et al., 2021), and that asymmetric training with smaller actors can be beneficial for real-world applications (Degrave et al., 2022).
2. We demonstrate that performance degradation and critic overfitting is largely due to poorer data collection, and this arises due to value underestimation.
   **Context:** This is somewhat surprising, as it stands in contrast to the *over-estimation* that's commonly addressed in many popular algorithms (Hasselt, 2010; Hasselt et al., 2016; Fujimoto et al., 2018). However, other papers have shown that underestimation can be an issue with the actor-critic algorithms that address overestimation (Ciosek et al., 2019; Li et al., 2023b; He & Hou, 2020).
3. We explore a number of approaches for mitigating the value underestimation and find the most effective one to be replacing the $\min$ term with an average or $\max$ term when combining the value estimates of two critics (as done in SAC).
   **Context:** Taking the minimum of two estimated $Q$-values will, by definition, be conservative; indeed, the idea was originally proposed to deal with over-estimation (Hasselt, 2010). Prior work has shown that resetting or regularizing the critic in particular improves plasticity (Ma et al., 2023; Nikishin et al., 2022; Liu et al., 2021) and can help mitigate value-estimation issues, particularly in the case of layer normalization (Nauman et al., 2024).

# PAC Apprenticeship Learning with Bayesian Active Inverse Reinforcement Learning

**Ondrej Bajgar, Dewi S.W. Gould, Jonathon Liu,**
**Alessandro Abate, Konstantinos Gatsis, Michael A. Osborne**

**Keywords:** inverse reinforcement learning, active learning, imitation learning, Bayesian methods

## Summary

As AI systems become increasingly autonomous, reliably aligning their decision-making to human preferences is essential. Inverse reinforcement learning (IRL) offers a promising approach to infer preferences from demonstrations. These preferences can then be used to produce an apprentice policy that performs well on the demonstrated task. However, in domains like autonomous driving or robotics, where errors can have serious consequences, we need not just good average performance but reliable policies with formal guarantees – yet obtaining sufficient human demonstrations for reliability guarantees can be costly. *Active* IRL addresses this challenge by strategically selecting the most informative scenarios for human demonstration. We introduce PAC-EIG, an information-theoretic acquisition function that directly targets probably-approximately-correct (PAC) guarantees for the learned policy – providing the first such theoretical guarantee for active IRL with noisy expert demonstrations. Our method maximises information gain about the regret of the apprentice policy, efficiently identifying states requiring further demonstration. We also present Reward-EIG as an alternative when learning the reward itself is the primary objective. Focusing on finite state-action spaces, we prove convergence bounds, illustrate failure modes of prior heuristic methods, and demonstrate our method's advantages experimentally.

## Contribution(s)

1. We formulate two principled information theoretic acquisition functions for active inverse reinforcement learning with Boltzmann rational demonstrations: Reward-EIG and PAC-EIG.
   **Context:** This gives a more principled alternative to previous, heuristic acquisition functions of Lopes et al. (2009), Brown et al. (2018), and Kweon et al. (2023).
2. For RegretEIG, we prove a lower bound on the expected number of steps of active learning needed to reach a probably-approximately-correct (PAC) apprentice policy.
   **Context:** This a first such proof for active IRL with an expert that is not perfectly rational. Metelli et al. (2021); Lindner et al. (2022) presented results for the, in many respects much simpler, case of perfectly optimal expert, focusing especially on transfer of a learnt reward to new environment dynamics.

# Average-DICE: Stationary Distribution Correction by Regression

**Fengdi Che, Bryan Chan, Chen Ma, A. Rupam Mahmood**

**Keywords:** Off-Policy Evaluation, State Distribution Correction, Importance Sampling, Distribution Shift.

## Summary

Off-policy policy evaluation (OPE), an essential component of reinforcement learning, has long suffered from stationary state distribution mismatch, undermining both stability and accuracy of OPE estimates. While existing methods correct distribution shifts by estimating density ratios, they often rely on expensive optimization or backward Bellman-based updates and struggle to outperform simpler baselines. We introduce Average-DICE, a computationally simple Monte Carlo estimator for the density ratio that averages discounted importance sampling ratios, providing an unbiased and consistent correction. Average-DICE extends naturally to nonlinear function approximation using regression, which we roughly tune and test on OPE tasks based on Mujoco Gym environments and compare with state-of-the-art density-ratio estimators using their reported hyperparameters. In our experiments, Average-DICE is at least as accurate as state-of-the-art estimators and sometimes offers orders-of-magnitude improvements. However, a sensitivity analysis shows that best-performing hyperparameters may vary substantially across different discount factors, so a re-tuning is suggested.

## Contribution(s)

1. We reformulate the state distribution ratio between the discounted stationary distribution of the target policy and the undiscounted stationary distribution of the behaviour policy as a new consistent estimator, leveraging a dataset collected under the behaviour policy.
2. We show that this consistent estimator corrects state distribution shifts in off-policy data, and reweighting each data point with our estimator provides an unbiased estimate for any function.
3. We introduce Average-DICE, an algorithm that estimates density ratios via regression.
4. We prove the convergence of our update rules under linear function approximation.
5. We evaluate Average-DICE against prior algorithms and demonstrate its fast convergence and low policy evaluation error.
   **Context:** Our algorithm is sensitive to changes in the environment, including the discount factor and the trajectory length.

# V-Max: A Reinforcement Learning Framework for Autonomous Driving

**Valentin Charraut**[1,†], **Waël Doulazmi**[1,2,†], **Thomas Tournaire**[1,†], **Thibault Buhet**[1]
{firstname.lastname}@valeo.com

1 **Valeo Brain**
2 **Centre for Robotics, Mines Paris - PSL**

† Equal contribution

## Abstract

Learning-based decision-making has the potential to enable generalizable Autonomous Driving (AD) policies, reducing the engineering overhead of rule-based approaches. Imitation Learning (IL) remains the dominant paradigm, benefiting from large-scale human demonstration datasets, but it suffers from inherent limitations such as distribution shift and imitation gaps. Reinforcement Learning (RL) presents a promising alternative, yet its adoption in AD remains limited due to the lack of standardized and efficient research frameworks. To this end, we introduce V-Max, an open research framework providing all the necessary tools to facilitate RL research for AD. V-Max is built on Waymax (Gulino et al., 2023), a hardware-accelerated AD simulator designed for large-scale experimentation. We extend it using ScenarioNet's (Li et al., 2023b) approach, enabling the fast simulation of diverse AD datasets. V-Max integrates a set of observation and reward functions, transformer-based encoders, and training pipelines. Additionally, it includes adversarial evaluation settings and an extensive set of evaluation metrics. Through a large-scale benchmark, we investigate how network architectures, observation functions, training data, and reward shaping impact RL performance. Code is available at: github.com/valeoai/v-max

## 1 Introduction

Reinforcement Learning (RL, Sutton & Barto (2018)) has proven to be a powerful approach for controlling real-world systems, with milestones in dexterous robotic manipulation and industrial process control (Rajeswaran et al., 2018; Degrave et al., 2022). RL's ability to learn adaptive policies through closed-loop interaction makes it an appealing framework for Autonomous Driving (AD, Kiran et al. (2022)), where decision-making agents must continuously respond to unseen scenarios and distribution shifts while maintaining high levels of robustness.

However, applying RL to real-world tasks such as AD introduces significant challenges, particularly regarding sample efficiency and training environments. As a result, RL remains underused in AD research due to practical constraints. Imitation Learning (IL, Bansal et al. (2019)) is often favored instead, as it capitalizes on vast driving datasets collected by vehicle fleets and reduces decision-making to a supervised learning task. The absence of RL-compatible environments made RL unusable in the nuPlan challenge (Karnchanachari et al., 2024), which led the organizers to conclude that learning-based methods could not compete with simple rule-based baselines (Dauner et al., 2023).

# Offline Action-Free Learning of Ex-BMDPs by Comparing Diverse Datasets

**Alexander Levine, Peter Stone, Amy Zhang**

**Keywords:** representation learning, action-free RL, Ex-BMDP, controllable state representations

## Summary

While sequential decision-making environments often involve high-dimensional observations, not all features of these observations are relevant for control. In particular, the observation space may capture factors of the environment which are not controllable by the agent, but which add complexity to the observation space. The need to ignore these "noise" features in order to operate in a tractably-small state space poses a challenge for efficient policy learning. Due to the abundance of video data available in many such environments, task-independent representation learning from action-free offline data offers an attractive solution. However, recent work has highlighted theoretical limitations in action-free learning under the Exogenous Block MDP (Ex-BMDP) model, where temporally-correlated noise features are present in the observations. To address these limitations, we identify a realistic setting where representation learning in Ex-BMDPs becomes tractable: when action-free video data from multiple agents with differing policies are available. Concretely, this paper introduces CRAFT (Comparison-based Representations from Action-Free Trajectories), a sample-efficient algorithm leveraging differences in controllable feature dynamics across agents to learn representations. We provide theoretical guarantees for CRAFT's performance and demonstrate its feasibility on a toy example, offering a foundation for practical methods in similar settings.

## Contribution(s)

1. We present a provably sample-efficient algorithm, CRAFT, that can learn high-accuracy latent state encoders under the Ex-BMDP model, when provided with two sets of offline observation trajectories, *without action labels*, that are collected by two agents with sufficiently-distinct policies.
   **Context:** Misra et al. (2024) has shown that efficient representation learning in Ex-BMDPs using a single offline dataset without action labels is in general *not* possible. This work therefore represents to our knowledge the first *positive* theoretical result for this problem. The Ex-BMDP model was introduced by Efroni et al. (2022), who propose a provably sample-efficient algorithm for *online* representation learning in this model. Efroni et al. (2022) assume *near*-deterministic latent-state dynamics, while we make a strict determinism assumption on latent-state dynamics. However, the negative result given by Misra et al. (2024) applies even to the full-determinism variant.
2. We prove the correctness of CRAFT and prove sample-complexity bounds.
   **Context:** None.
3. We demonstrate the feasibility of CRAFT on a toy problem, and present the results.
   **Context:** None.

# One Goal, Many Challenges: Robust Preference Optimization Amid Content-Aware, Multi-Source Noise

**Amirabbas Afzali, Amirhossein Afsharrad, Seyed Shahabeddin Mousavi, Sanjay Lall**

**Keywords:** Reinforcement Learning from Human Feedback, Preference Optimization, Content-Aware Noise, Robust Preference Learning

## Summary

Large Language Models (LLMs) have significantly advanced in generating human-like responses, largely due to Reinforcement Learning from Human Feedback (RLHF). However, RLHF methods often assume unbiased human annotations, which is rarely the case in real-world settings. This paper introduces Content-Aware Noise-Resilient Preference Optimization (CNRPO), a novel framework that explicitly models and mitigates content-dependent noise in preference learning. CNRPO employs a multi-objective optimization approach to disentangle true preferences from biased signals, improving robustness against multi-source annotation noise. Furthermore, we leverage backdoor attack mechanisms to efficiently identify, learn, and control bias-inducing triggers within a single model. Our theoretical analysis and extensive experiments on different synthetic noisy datasets demonstrate that CNRPO significantly enhances preference optimization in RLHF by aligning models with primary human preferences while controlling for secondary noise factors, such as response length and harmfulness.

## Contribution(s)

1. We introduce Content-Aware Noise-Resilient Preference Optimization (CNRPO), a framework that explicitly models content-dependent noise in preference learning.
   **Context:** Prior work on preference optimization has addressed noise in annotations but has not explicitly accounted for content-aware biases (Chowdhury et al., 2024; Gao et al., 2024).
2. We leverage multi-objective optimization to disentangle and control noise sources, enabling more robust preference learning.
   **Context:** Existing approaches typically assume uniform noise distributions, which fail to capture the complexity of multi-source biases in preference datasets (Mitchell, 2023; Liang et al., 2024).
3. We incorporate backdoor attack mechanisms as a novel tool to understand and mitigate biases in preference annotations.
   **Context:** Backdoor attacks have been explored in adversarial settings (Pathmanathan et al., 2024), but their use in bias control for preference learning is a new contribution.
4. We provide theoretical analysis and extensive empirical validation on different synthetic noisy datasets, demonstrating the effectiveness of CNRPO in mitigating biases.
   **Context:** Prior studies have evaluated preference learning under noise but lack theoretical guarantees and controlled empirical validation across multiple bias sources.

# A Timer-Based Hybrid Supervisor for Robust, Chatter-Free Policy Switching

**Jan de Priester, Ricardo G. Sanfelice**

**Keywords:** Chattering, robustness, value function-based switching, hybrid control.

## Summary

We address the challenge of switching among multiple learned policies in reinforcement learning control systems, where conventional value function–based methods can lead to chattering in the presence of small measurement noise. Our goal is to design a switching logic that assures asymptotic stability and maintains a robustness margin so that rapid switching is prevented under bounded measurement noise. To this end, we propose a timer-based hybrid supervisor that integrates a resettable timer that enforces a minimum dwell time on the active policy. This dwell time is adaptively adjusted by predicting the evolution of the state of the system, ensuring that a switch occurs only when a significantly better alternative is predicted. We derive sufficient conditions under which the hybrid supervisor is guaranteed to exhibit non-Zeno behavior and render a compact set robustly globally asymptotically stable in the presence of bounded measurement noise. Simulation results on representative decision-making problems demonstrate that our hybrid supervisor is robust under noisy conditions where a conventional switching strategy fails.

## Contribution(s)

1. This paper presents a hybrid supervisor that maintains the asymptotic stability properties of the underlying policies and prevents chattering between the policies under bounded measurement noise. The hybrid supervisor deploys a timer-based mechanism to predict and enforce a dwell period between policy switches. Sufficient conditions are presented under which the hybrid supervisor is guaranteed to exhibit non-Zeno behavior and render a compact set robustly globally asymptotically stable in the presence of bounded measurement noise.
   **Context:** Chattering refers to the phenomenon of a system rapidly switching its decision due to measurement noise that results in inefficient or destabilizing behavior. Existing chattering-mitigation strategies in RL rely on partitioning the state space into overlapping regions to define switching conditions, with the overlap situated where chattering is observed. Defining these overlapping regions necessitates some insight into their expected location, making this approach more suitable when such prior knowledge is available. Alternative timer-based approaches that *may* prevent chattering impose a fixed lower bound on the time between switches. These methods are not designed to address chattering under measurement noise and thus lack formal guarantees that bounded disturbances will not induce chattering.

# Deep Reinforcement Learning with Gradient Eligibility Traces

**Esraa Elelimy, Brett Daley, Andrew Patterson, Marlos C. Machado, Adam White, Martha White**

**Keywords:** Deep RL, Gradient TD, Eligibility Traces, PPO

## Summary

Achieving fast and stable off-policy learning in deep reinforcement learning (RL) is challenging. Most existing methods rely on semi-gradient temporal-difference (TD) methods for their simplicity and efficiency, but are consequently susceptible to divergence. While more principled approaches like Gradient TD (GTD) methods have strong convergence guarantees, they have rarely been used in deep RL. Recent work introduced the generalized Projected Bellman Error ($\overline{\text{PBE}}$), enabling GTD methods to work efficiently with nonlinear function approximation. However, this work is limited to one-step methods, which are slow at credit assignment and require a large number of samples. In this paper, we extend the generalized $\overline{\text{PBE}}$ objective to support multistep credit assignment based on the $\lambda$-return and derive three gradient-based methods that optimize this new objective. We provide both a forward-view formulation compatible with experience replay and a backward-view formulation compatible with streaming algorithms. Finally, we evaluate the proposed algorithms and show that they outperform both PPO and StreamQ in MuJoCo and MinAtar environments, respectively. [a]

[a] Code available at https://github.com/esraaelelimy/gtd_algos

## Contribution(s)

1. We extend the generalized $\overline{\text{PBE}}$ to incorporate multistep credit assignment based on $\lambda$-returns, defining a new objective, the $\overline{\text{PBE}}(\lambda)$ (Section 3).
   **Context:** Patterson et al. (2022) introduced the generalized $\overline{\text{PBE}}$, which unifies and generalizes previously known objectives for value estimation. However, it was only defined for the one-step TD error.
2. We derive three Gradient TD algorithms that optimize our proposed objective. We derive both the forward view with the $\lambda$-return (Section 4) and the backward view with eligibility traces (Section 6).
   **Context:** Gradient TD methods were originally introduced with linear function approximation (Sutton et al., 2009), with a limited extension to nonlinear function approximation that required second-order information (Maei et al., 2009). The recent work by Patterson et al. (2022) extended these methods to non-linear function approximation without a need for second-order information. However, it was limited to the one-step TD error.
3. We introduce Gradient PPO, a policy gradient algorithm that uses our sound forward-view value estimation algorithms (Section 5).
   **Context:** PPO (Schulman et al., 2017) is a widely-used policy gradient method that relies on semi-gradient TD updates for value estimation. We build on PPO by replacing the value estimation component with a new one that uses Gradient TD methods. This change required non-trivial modification to PPO, resulting in our new algorithm, Gradient PPO. Gradient PPO is the first policy gradient method that uses Gradient TD algorithms in a deep RL setting with a replay buffer.
4. We introduce QRC($\lambda$), which uses our backward-view eligibility traces and is suitable for streaming settings. (Section 6).
   **Context:** Backward-view algorithms can make updates on each time step without delay, making them efficient in streaming settings (Elsayed et al., 2024). QRC($\lambda$) is the first backward-view algorithm that uses Gradient TD methods in the streaming deep RL.

# On Slowly-Varying Non-Stationary Bandits

**Ramakrishnan Krishnamurthy , Aditya Gopalan**

**Keywords:** Multi-armed bandits, Slowly-varying Non-stationary rewards, Change point detection.

## Summary

We consider minimisation of dynamic regret in non-stationary multi-armed bandits with a slowly varying property. Namely, we assume that arms' rewards are stochastic and independent over time, but that the absolute difference between the expected rewards of any arm at any two consecutive time-steps is at most a drift limit $\delta > 0$. For this setting that has not received enough attention in the past, we give a new algorithm and establish the first instance-dependent regret upper bound for slowly varying non-stationary bandits. The analysis, in turn, relies on a novel characterization of the instance as a *detectable gap* profile that depends on the expected arm reward differences. We also provide the first minimax regret lower bound for this problem, enabling us to show that our algorithm is essentially minimax optimal. Also, this lower bound we obtain establishes that the seemingly easier slowly-varying bandits problem is at least as hard as the more general total variation-budgeted bandits problem in the minimax sense. We complement our theoretical results with experimental illustrations.

## Contribution(s)

1. We design a new algorithm for the problem of slowly-varying non-stationary multi-armed bandits. We show an instance-dependent regret upper bound for this algorithm. For this, we come up with a novel instance-dependent quantity that we call 'detectable gap'.
   **Context:** To the best of our knowledge, this is the first instance-dependent regret bound for the slowly-varying settings. Instance-dependent bounds have so far been elusive in any continuously varying bandits (both total variation-budgeted setting and slowly-varying setting).
2. We show a minimax regret upper bound for our algorithm and establish that it is minimax optimal.
   **Context:** Besbes et al. (2014) already show a minimax optimal algorithm for the more general total variation-budgeted bandits problem, so ours is not the first/only minimax optimal algorithm.
3. We show a fundamental lower bound for slowly-varying non-stationary bandits problem. This bound matches our upper bound and also matches the known lower bound of the total variation-budgeted bandits problem. This establishes that the more constrained slowly-varying setting is at least as hard (in a worst case sense) as the more general total variation-budgeted setting.
   **Context:** To the best of our knowledge, this is the first lower bound for the slowly-varying non-stationary bandits problem.
4. We experimentally evaluate the performance of our new algorithm, and compare with existing approaches from the literature.
   **Context:** None.

# Focused Skill Discovery: Learning to Control Specific State Variables while Minimizing Side Effects

**Jonathan Colaço Carr, Qinyi Sun, Cameron Allen**

**Keywords:** Skill Discovery, Hierarchical Reinforcement Learning

## Summary

Skills are essential for unlocking higher levels of problem solving. A common approach to discovering these skills is to learn ones that reliably reach different states, thus empowering the agent to control its environment. However, existing skill discovery algorithms often overlook the natural state variables present in many reinforcement learning problems, meaning that the discovered skills lack control of specific state variables. This can significantly hamper exploration efficiency, make skills more challenging to learn with, and lead to negative side effects in downstream tasks when the goal is under-specified. We introduce a general method that enables these skill discovery algorithms to learn *focused skills*—skills that target and control specific state variables. Our approach improves state space coverage by a factor of three, unlocks new learning capabilities, and automatically avoids negative side effects in downstream tasks.

## Contribution(s)

1. This paper presents a general method for modifying existing skill discovery algorithms such that the learned skills control individual state variables.
   **Context:** Prior work by Hu et al. (2024) explores a similar method, but is limited to skills that are discovered by maximizing mutual information. Our method is compatible with a wider variety of skill discovery methods, which we demonstrate with Lipschitz-constrained Skill Discovery (LSD) (Park et al., 2022), as well as with two mutual-information driven skill discovery methods (Gregor et al., 2017; Eysenbach et al., 2018).
2. We show that applying our algorithm to three skill discovery algorithms improves exploration efficiency by a factor of three as compared to their original variations.
   **Context:** We assess exploration efficiency by measuring the fraction of unique final states that can be reached after executing all possible skill-chain combinations of a given length, and then calculating the area under the curve for different skill-chain lengths.
3. We show that our approach can learn skills that automatically avoid unwanted side effects when the goal is underspecified (i.e. when the goal is only specified explicitly for a subset of the state variables, but changes to the others should still be minimized).
   **Context:** This is an important area of research for improving the safety and effectiveness of learned skills. Prior work by Turner et al. (2020) and Krakovna et al. (2020) has discussed underspecified objectives but has not explored how skill discovery as a pre-training step can help mitigate these effects.
4. Compared to an existing method, we show that our approach learns skills that can be significantly more effective in downstream tasks where side effects should be minimized.
   **Context:** We measure the effectiveness in downstream tasks using the number of steps to reach the goal vs. training episode. The method we compare against (Hu et al., 2024) aims to learn skills that control individual state variables while minimizing side effects, but is less effective than our method on two tasks, both in terms of final performance and learning efficiency.

# Goals vs. Rewards: A Preliminary Comparative Study of Objective Specification Mechanisms

**Septia Rani, Serena Booth, Sarath Sreedharan**

**Keywords:** objective specification, goals, rewards.

## Summary

This paper studies two popular objective specification mechanisms for sequential decision-making problems: goals and rewards. We investigate how easy it is for non-AI experts to use these different specification mechanisms effectively. Namely, we investigate how effectively people can use these mechanisms to (a) correctly direct an AI system or robot to generate some desired behavior and (b) predict the behavior encoded in a given objective specification. We perform a user study to assess these questions. In addition, we present a formalization of the problems of objective specification and behavior prediction, and we characterize *underspecification* and *overspecification*. While participants have a strong preference for using goals as an objective specification mechanism, we find a surprising result: even non-expert users are equally capable of specifying and interpreting reward functions.

## Contribution(s)

1. The paper assesses how well non-expert users can effectively make use of goal and reward specification mechanisms. In particular, we study whether they (a) can use these mechanisms to generate specifications that result in some intended target behavior and (b) can predict behavior that could result from the given specification.
   **Context:** We are unaware of any works that perform such human-centric comparisons. The closest works we know of focus purely on how successful engineers are in hand-crafting reward functions (cf. (Knox et al., 2023; Booth et al., 2023)).
2. We provide a formal definition of the specification and prediction task to support comparisons between reward functions and goals. We also provide a formal characterization of the conditions under which an objective can be said to be overspecified or underspecified.
   **Context:** While there are existing works that have tried to model objective misspecification (e.g., Mechergui & Sreedharan (2024)), underspecification (e.g., Shah et al. (2022)), and misspecification (e.g., Amodei et al. (2016)), these definitions have not been formalized to cover and compare multiple specification modalities.
3. Our results present evidence that the non-expert users' ability to correctly specify and interpret reward functions is comparable to their ability to provide goal specifications. However, we see a clear difference in their preferences between the two metrics: they overwhelmingly prefer the goal mechanism.
   **Context:** We are unaware of any prior works that point to parity in user ability to leverage the two objective specification mechanisms. This result may imply that developing novel interfaces for reward functions could help users of RL techniques to utilize reward functions more effectively. Mechanisms like reward machines are one such promising mechanism (Icarte et al., 2022).

# An Analysis of Action-Value Temporal-Difference Methods That Learn State Values

**Brett Daley, Prabhat Nagarajan, Martha White, Marlos C. Machado**

**Keywords:** TD learning, QV-learning, Dueling DQN, advantage estimation.

## Summary

The hallmark feature of temporal-difference (TD) learning is bootstrapping: using value predictions to generate new value predictions. The vast majority of TD methods for control learn a policy by bootstrapping from a single action-value function (e.g., Q-learning and Sarsa). Significantly less attention has been given to methods that bootstrap from two asymmetric value functions: i.e., methods that learn state values as an intermediate step in learning action values. Existing algorithms in this vein can be categorized as either QV-learning or AV-learning. Though these algorithms have been investigated to some degree in prior work, it remains unclear if and when it is advantageous to learn two value functions instead of just one—and whether such approaches are theoretically sound in general. In this paper, we analyze these algorithmic families in terms of convergence and sample efficiency. We find that while both families are more efficient than Expected Sarsa in the prediction setting, only AV-learning methods offer any major benefit over Q-learning in the control setting. Finally, we introduce a new AV-learning algorithm called Regularized Dueling Q-learning (RDQ), which significantly outperforms Dueling DQN in the MinAtar benchmark.

## Contribution(s)

1. We prove the expected contraction of QV-learning for on-policy prediction.
   **Context:** Wiering (2005) introduced the QV-learning algorithm, but omitted a convergence proof. To our knowledge, there is no published convergence proof of QV-learning to date.
2. We raise the issue that QVMAX, the main off-policy control variant of QV-learning, is biased. We empirically demonstrate that such bias can significantly impact performance. We then introduce a new, unbiased algorithm, BC-QVMAX, and empirically demonstrate that it converges to a similar solution to that of Q-learning in our considered settings.
   **Context:** Wiering & van Hasselt (2009) introduced QVMAX without theoretical justification, heuristically mirroring the Q-learning update. Its bias has not been identified until now. Our empirical results were obtained with parametric MDP experiments that we designed; they should be interpreted only as anecdotal evidence for the convergence of BC-QVMAX.
3. In the context of AV-learning algorithms, we formalize a tabular version of Dueling DQN, and we introduce a new algorithm, Regularized Dueling Q-learning (RDQ). RDQ addresses the identifiability issue of the naive dueling decomposition by using an $l_2$ penalty instead of subtracting the mean advantage. We empirically demonstrate that, given the same network architecture, RDQ significantly outperforms Dueling DQN in the MinAtar domain.
   **Context:** Wang et al. (2016) introduced Dueling DQN from the perspective of an improvement to the neural network architecture used by DQN. RDQ is the result of relaxing the semantics behind estimating $Q(s, a)$ through two value functions. Instead of generating $Q(s, a)$ from approximations of $V(s)$ and $A(s, a)$, RDQ searches for the closest point on the hyperplane defined by two arbitrary functions that sum to $Q(s, a)$. We used tuned versions of DQN and Dueling DQN as baselines for MinAtar (Obando-Ceron & Castro, 2021).

# Effect of a slowdown correlated to the current state of the environment on an asynchronous learning architecture

**Idriss Abdallah, Laurent Ciarletta , Patrick Hénaff, Matthieu Bonavent, Jonathan Champagne**

**Keywords:** Deep reinforcement learning, Asynchronous architecture, Environment slowdown

## Summary

In an industrial context, we apply deep reinforcement learning (DRL) to a simulator of an unmanned underwater vehicle (UUV). This UUV is moving in a complex environment that needs to compute acoustic propagation in very different scenarios. Consequently, the computation time per timestep varies greatly due to the complexity of the acoustic situation and the variation in the number of elements simulated. Therefore, we use an asynchronous actor-learner parallelization scheme to avoid any loss of computational resource efficiency. However, there is a strong correlation between the current state of the environment and this variability in computation time. The classical benchmarks in the DRL are not representative of our environment slowdowns, neither in magnitude nor in their correlation with the current observation. The aim of this paper is therefore to investigate the possible existence of a bias that could be induced by an observation-correlated slowdown in the case of a DRL algorithm using an asynchronous architecture. We empirically demonstrate the existence of such a bias in a modified Cartpole environment. We then study the evolution of this bias as a function of several parameters: the number of parallel environments, the exploration, and the positioning of slowdowns. Results reveal that the bias is highly dependent on the capacity of the policy to discover trajectories that avoid the slowdown areas.

## Contribution(s)

1. We show that classical reinforcement learning benchmarks are not representative of our industrial environment in terms of the effects of slowdowns correlated with observation.
   **Context:** We based our analysis on the two most used deep reinforcement learning benchmarks : Mujoco (Todorov et al., 2012) and Atari (Bellemare et al., 2012) using the learning framework TorchRL (Bou et al., 2024).
2. We provide empirical evidence on a modified version of the Cartpole environment that an environment with observation-correlated slowdowns can induce a bias on the data generated and the learned policy for an algorithm using an asynchronous architecture.
   **Context:** We used the Dueling Double Deep Q-Learning (Wang et al., 2016) with the actor-learner architecture described by Espeholt et al. (2018) which decouples threads when generating transitions to reduce inter-process synchronization to achieve greater scalability.
3. We investigated bias changes as a function of the number of parallel actors, the exploration, and the positioning of the slowdown zone. Our results show that the bias is highly dependent on the capacity of the policy to find trajectories that avoid the slowdown areas.
   **Context:** None

# PEnGUiN: Partially Equivariant Graph NeUral Networks for Sample Efficient MARL

**Joshua McClellan, Greyson Brothers, Furong Huang, Pratap Tokekar**

**Keywords:** sample efficiency, reinforcement learning, symmetry, equivariance, geometric guarantees, inductive bias

## Summary

Equivariant Graph Neural Networks (EGNNs) excel at Multi-Agent Reinforcement Learning (MARL) problems by harnessing symmetries in observations, but struggle in real-world environments where symmetries may be broken to varying degrees. We introduce *Partially Equivariant Graph Neural Networks (PEnGUiN)*, a novel architecture that learns to exploit partial symmetries. PEnGUiN blends equivariant and non-equivariant updates via a learnable parameter, adapting to the degree and type of symmetry present and bridging the gap between fully equivariant and non-equivariant models. In addition, we formalize types of partial equivariance common to real-world environments (subgroup, feature-wise, subspace, and approximate). Experiments on MARL benchmarks demonstrate PEnGUiN's superior performance and robustness compared to EGNNs and GNNs in asymmetric settings. PEnGUiN learns where equivariance holds, improving applicability to real-world MARL problems.

## Contribution(s)

1. We present the first generalization of Equivariant Graph Neural Networks (EGNN) to Partial Equivariance with our novel neural network architecture Partially Equivariant Graph Neural Networks (PEnGUiN). We show theoretically that PEnGUiN unifies fully equivariant (EGNN) and non-equivariant (GNN) representations within a single architecture, controlled by a learnable parameter called the symmetry score.
   **Context:** PEnGUiN builds on EGNN (Satorras et al., 2021) and E2GN2 (McClellan et al., 2024), and is designed to handle environments with asymmetries, unlike prior work that primarily focuses on full equivariance.
2. We show the first Partially Equivariant Neural Network applied to Multi-Agent Reinforcement Learning, leading to improved performance over GNNs and EGNNs in MARL.
   **Context:** Prior work has applied equivariance to MARL (Pol et al., 2021; McClellan et al., 2024), these approaches typically assume full equivariance.
3. We formally define and categorize several types of partial equivariance relevant to Multi-Agent Reinforcement Learning (MARL), including subgroup equivariance, feature-wise equivariance, subspace equivariance, and approximate equivariance.
   **Context:** While specific instances of broken symmetries have been discussed (Chen et al., 2023; Park et al., 2024), our work provides a unified and comprehensive categorization tailored to MARL.
4. Through experiments on Multi-Particle Environments (MPE) and the highway-env benchmark, we empirically validate that PEnGUiN outperforms both EGNNs and standard GNNs in MARL tasks with various types of asymmetries.
   **Context:** None

# Shaping Laser Pulses with Reinforcement Learning

**Francesco Capuano, Davorin Peceli, Gabriele Tiboni**

**Keywords:** Applied RL, DRL for Science, Sim-to-real, Domain Randomization, Ultra-short pulses

## Summary

High Power Laser (HPL) systems operate in the attoseconds regime—the shortest timescale ever created by humanity. HPL systems are instrumental in high-energy physics, leveraging ultra-short impulse durations to yield extremely high intensities, which are essential for both practical applications and theoretical advancements in light-matter interactions. Traditionally, the parameters regulating HPL optical performance have been manually tuned by human experts, or optimized using black-box methods that can be computationally demanding. Critically, black box methods rely on stationarity assumptions overlooking complex dynamics in high-energy physics and day-to-day changes in real-world experimental settings, and thus need to be often restarted. Deep Reinforcement Learning (DRL) offers a promising alternative by enabling sequential decision making in non-static settings. This work explores the feasibility of applying DRL to HPL systems, extending the current research by (1) learning a control policy relying solely on non-destructive image observations obtained from readily available diagnostic devices, and (2) retaining performance when the underlying dynamics vary. We evaluate our method across various test dynamics, and observe that DRL effectively enables cross-domain adaptability, coping with dynamics' fluctuations while achieving 90% of the target intensity in test environments.

## Contribution(s)

1. We demonstrate the benefits of using Deep Reinforcement Learning to optimize High Power Laser systems over currently dominant approaches based on gradient-free optimization.
   **Context:** Prior works on laser optimization focused on black-box optimization techniques which assume stationarity, require costly real-world function evaluations, and can endanger the system at test-time.
2. We learn a control policy directly from images, which are made available via widespread diagnostics devices.
   **Context:** Instead of relying on noisy and lengthy processes to obtain structured representations of the system's state, we leverage unstructured observations coming from diagnostic devices as inputs for the control policy.
3. We train control policies entirely in simulation and successfully transfer them across unknown, varying dynamics, showing robustness to different parametrizations.
   **Context:** Transferring policies is hindered by the discrepancies across different domains, and Domain Randomization, a promising technique widely explored in then field of robot learing, can be leveraged to ensure robustness.

# Cascade - A sequential ensemble method for continuous control tasks

**Schmöcker R. , Dockhorn A.**

**Keywords:** Ensemble learning, reinforcement learning, continuous control, PPO.

## Summary

Though reinforcement learning has been successfully applied to a variety of domains, there is still room left for improvement, in particular, in terms of the final performance. Ensemble Reinforcement Learning (ERL) tries to enhance reinforcement learning techniques by using multiple models or algorithms. We propose a novel ERL technique, called Cascade which in the context of continuous control tasks and with PPO as the base training algorithm clearly outperforms standard PPO in terms of the final performance. To shine light on the working mechanisms of Cascade, we conduct ablation studies, showing how the different components of Cascade contribute to its overall performance. Furthermore, we demonstrate that Cascade has a robust monotonicity as the ensemble's performance increases with each additional base agent even when weak base agents are added in large numbers.

## Contribution(s)

1. The proposition of a novel Ensemble Reinforcement Learning (ERL) algorithm Cascade for continuous control tasks that outperforms its base learner when using PPO as the underlying reinforcement learning algorithm.
   **Context:** To the best of our knowledge, there is no prior work where the ensemble policy uses a convex combination of its base learners and still gains a significant performance advantage.
2. By multiple ablation studies, we investigate the mechanisms contributing to Cascade's performance.
   **Context:** We show that Cascade relies on all base learners being trained at all stages of the training process as well as Cascade relying on sequentially adding base learners instead of starting with the final network. Lastly, we show that Cascade can chain an arbitrary number of base learners of arbitrary strengths without a loss in performance.

# Reinforcement Learning with Adaptive Temporal Discounting

**Sahaj Singh Maini, Zoran Tiganj**

**Keywords:** Adaptive Temporal Discounting, Weber-Fechner law, Log-compressed Timeline.

## Summary

Conventional reinforcement learning (RL) methods often fix a single discount factor for future rewards, limiting their ability to handle diverse temporal requirements. We propose a framework that utilizes an interpretation of the value function as a Laplace transform. By training an agent across a spectrum of discount factors and applying an inverse transform, we recover a log-compressed representation of expected future reward. This representation enables post hoc adjustments to the discount function (e.g., exponential, hyperbolic, or finite horizon) without retraining. Furthermore, by precomputing a library of policies, the agent can dynamically select the policy that maximizes a newly specified discount objective at runtime, effectively constructing a hybrid policy to handle varying temporal objectives. The properties of this log-compressed timeline are consistent with human temporal perception as described by the Weber-Fechner law, theoretically enhancing efficiency in scale-free environments by maintaining uniform relative precision across timescales. We demonstrate this framework in a grid-world navigation task where the agent adapts to different time horizons.

## Contribution(s)

1. We extend prior methods for computing log-compressed representations of expected future reward to a dynamic policy evaluation setting, showing that when a library of policies is available, this representation enables immediate re-evaluation of these policies under any desired discount function.
   **Context:** Prior work established methods for computing log-compressed representations of expected future reward (Momennejad & Howard, 2018; Tiganj et al., 2019; Tano et al., 2020; Masset et al., 2023). Our extension focuses on dynamical policy evaluation with arbitrary temporal discounting. We evaluate the approach under idealized conditions where true value functions are accessible, demonstrated through a grid-world navigation task.
2. Leveraging the Weber–Fechner link between log-time codes and human perception, we analytically show that log-compressed representations enable efficient decision-making in scale-free environments by maintaining uniform relative precision across timescales.
   **Context:** The Weber-Fechner law Fechner (1860/1912) is a widely referenced principle in psychophysics, stating that perceived magnitude is proportional to the logarithm of stimulus intensity, implying a logarithmic scale.

# Average-Reward Soft Actor-Critic

**Jacob Adamczyk, Volodymyr Makarenko,**
**Stas Tiomkin, Rahul V. Kulkarni**

**Keywords:** average-reward, MaxEnt, entropy-regularization, actor-critic, deep RL.

## Summary

The average-reward formulation of reinforcement learning (RL) has drawn increased interest in recent years for its ability to solve temporally-extended problems without relying on discounting. Meanwhile, in the discounted setting, algorithms with entropy regularization have been developed, leading to improvements over deterministic methods. Despite the distinct benefits of these approaches, deep RL algorithms for the entropy-regularized average-reward objective have not been developed. While policy-gradient based approaches have recently been presented for the average-reward literature, the corresponding actor-critic framework remains less explored. In this paper, we introduce an average-reward soft actor-critic algorithm to address these gaps in the field. We validate our method by comparing with existing average-reward algorithms on standard RL benchmarks, achieving superior performance for the average-reward criterion.

## Contribution(s)

1. We generalize the soft actor-critic (SAC) algorithm from the discounted to the average-reward setting.
   **Context:** Haarnoja et al. (2018b) derived a MaxEnt RL algorithm, soft actor-critic, for the discounted setting. We derive theoretical results and implement new algorithmic techniques to adapt SAC to the average-reward setting.
2. We extend the policy improvement theorem to the entropy-regularized average-reward objective.
   **Context:** Previous work demonstrated the policy improvement theorem separately in discounted MaxEnt RL (Haarnoja et al., 2018b) and average-reward (un-regularized) RL (Zhang & Tan, 2024). We close this gap by analyzing the theoretical properties of policy improvement in the entropy-regularized average-reward setting.
3. We experimentally demonstrate the advantage of our approach against available baselines in standard control environments.
   **Context:** We compare our algorithm with existing baseline average-reward methods: ARO-DDPG (Saxena et al., 2023), ATRPO (Zhang & Ross, 2021), and APO (Ma et al., 2021).

# Human-Level Competitive Pokémon via Scalable Offline Reinforcement Learning with Transformers

**Jake Grigsby, Yuqi Xie[†], Justin Sasek[†], Steven Zheng[†], Yuke Zhu**

**Keywords:** Pokémon , Offline RL, Imitation Learning

## Summary

Competitive Pokémon Singles (CPS) is a popular strategy game where players learn to exploit their opponent based on imperfect information in battles that can last more than one hundred stochastic turns. AI research in CPS has been led by heuristic tree search and online self-play, but the game may also create a platform to study adaptive policies trained offline on large datasets. We develop a pipeline to reconstruct the first-person perspective of an agent from logs saved from the third-person perspective of a spectator, thereby unlocking a dataset of real human battles spanning more than a decade that grows larger every day. This dataset enables a black-box approach where we train large sequence models to adapt to their opponent based solely on their input trajectory while selecting moves without explicit search of any kind. We study a progression from imitation learning to offline RL and offline fine-tuning on self-play data in the hardcore competitive setting of Pokémon's four oldest (and most partially observed) game generations. The resulting agents outperform a recent LLM Agent approach and a strong heuristic search engine. While playing anonymously in online battles against humans, our best agents climb to rankings inside the top $10\%$ of active players. All agent checkpoints, training details, datasets, and baselines are available at `metamon.tech`.

## Contribution(s)

1. We build and release an offline RL dataset comprising $3.5$M trajectories reconstructed from years of human gameplay in the complex decision-making task of Competitive Pokémon.
   **Context:** PokéChamp Karten et al. (2025) concurrently released a dataset of Pokémon battles. The datasets differ in that:
   - Ours covers all available data ($2014$-Present) for a smaller list of popular game modes. This provides more demonstrations per mode and explores the challenges of learning from strategies that evolve over time.
   - Ours is distributed in a flexible RL format that allows for customization of observations, actions, and rewards outside of LLM prompts.
   - Ours reconstructs the agent's partially observed perspective from spectator data with more accuracy thanks to a custom state-tracking and prediction pipeline designed for this purpose. Further discussion is provided in Appendix D and in our open-source release.
2. We demonstrate our dataset's ability to produce sequence policies that play Competitive Pokémon at a human level.
   **Context:** Prior work has used online self-play and heuristic search to build successful Pokémon agents in other rulesets.

# Adaptive Submodular Policy Optimization

**Branislav Kveton, Anup Rao, Viet Lai, Nikos Vlassis, David Arbour**

**Keywords:** policy gradients, submodularity, adaptive submodularity

## Summary

We propose KL-regularized policy optimization for adaptive submodular maximization, which is a framework for decision making under uncertainty with submodular rewards. Policy optimization of adaptive submodular functions justifies a surprisingly simple and efficient policy gradient update, where the optimized action only affects its immediate reward but not the future ones. It also allows us to learn adaptive submodular policies with large action spaces, such as those represented by large language models (LLMs). We prove that our policies monotonically improve as the regularization diminishes and converge to the optimal greedy policy. Our experiments show major gains in statistical efficiency, in both synthetic problems and LLMs.

## Contribution(s)

1. We propose KL-regularized policy optimization for adaptive submodular maximization.
   **Context:** There are prior works on gradient-based optimization of submodular (not adaptive) functions. See Paragraph 2 in Section 6. There are prior works on policy gradients in more general settings. See Paragraphs 1 and 3 in Section 6.
2. We derive more efficient policy gradient estimators than in more general settings, with $O(n)$ terms as opposing to $O(n^2)$, where $n$ is the horizon.
   **Context:** None
3. We prove that our policy converges to the optimal greedy policy for adaptive submodular maximization as the regularization diminishes (Theorem 1). We prove that our policies monotonically improve over reference policies used for their regularization as the regularization diminishes (Theorem 4).
   **Context:** None
4. We demonstrate the efficiency of new policy gradient estimators empirically, in both synthetic problems and LLMs (Section 5).
   **Context:** None

# Learning Fair Pareto-Optimal Policies in Multi-Objective Reinforcement Learning

**Umer Siddique, Peilang Li, Yongcan Cao**

**Keywords:** Multi-objective reinforcement learning, Deep reinforcement learning, Fair optimization, Welfare functions

## Summary

Fairness is important in multi-objective reinforcement learning (MORL), where policies must balance optimality and equity across objectives. While *single-policy* MORL methods can learn fair policies for fixed user preferences using welfare, they fail to generalize for different user preferences. To address this limitation, we propose a novel framework for fairness in *multi-policy* MORL, which learns a set of fair policies. Our theoretical analysis establishes that for concave and piecewise-linear welfare functions, fair policies remain in the convex coverage set (CCS). Additionally, we demonstrate that non-stationary and stochastic policies improve fairness over stationary and deterministic policies. Building on our theoretical analysis, we introduce three scalable methods: an extension of Envelope for fair stationary policies, a non-stationary counterpart using state-augmented accrued rewards, and a novel extension for learning stochastic policies. We validate our methods through extensive experiments across three domains and show that our methods fairer solutions as compared to MORL baselines.

## Contribution(s)

1. We introduce a novel framework for fairness in multi-policy MORL, which enables learning a set of fair policies for varying user preferences.
   **Context:** Prior work on fairness in MORL has mainly focused on a single policy for predefined preference weights via some welfare functions. Our framework generalizes fairness across multiple policies, which allow end users to select any policy provided by their preference weights.
2. We provide theoretical analysis demonstrating that for concave, piecewise-linear welfare functions, fair policies remain in the convex coverage set (CCS). Additionally, we establish that non-stationary and stochastic policies can enhance fairness over stationary and deterministic policies, respectively.
   **Context:** Existing work has explored fairness in RL for predefined preference weights but has not theoretically analyzed how non-stationary and stochastic policies can improve fairness for varying preference weights.
3. We propose three scalable methods for learning fair policies in MORL using a single parameterized network: (i) an extension of Envelope (Yang et al., 2019) for learning fair policies, (ii) a non-stationary extension that incorporates state-augmented accrued rewards to adaptively improve fairness, and (iii) a novel stochastic policy learning method that further enhances fairness.
   **Context:** Unlike prior work on MORL, which typically learns Pareto optimal policies, our methods efficiently learn a set of fair policies while maintaining scalability.

# Representation Learning and Skill Discovery with Empowerment

**Andrew Levy, Alessandro Allievi, George Konidaris**

**Keywords:** Unsupervised Skill Discovery, Representation Learning, Empowerment

## Summary

Representation learning and unsupervised skill discovery remain key challenges for training reinforcement learning agents. We show that the empowerment objective enables agents to simultaneously perform both representation learning and unsupervised skill discovery. Our theoretical analysis shows that empowerment provides a principled objective for learning sufficient statistic representations of observations. To jointly learn representations and skills, we use a tighter variational lower bound on mutual information relative to prior work, and we maximize this objective using a new actor-critic architecture. We also show empirically in a variety of settings that our approach enables agents to jointly learn representations and large skillsets conditioned on those representations.

## Contribution(s)

1. We prove that for any encoder that maps observations to a learned representation, the average empowerment achieved by the encoder is upper bounded by the average empowerment achieved by an encoder that outputs sufficient statistics of observations.
   **Context:** Prior work has proven that the average empowerment produced by an observation encoder is upper bounded by the average empowerment conditioned on the state representation (Capdepuy, 2011). We prove this is a looser upper bound than our own. This bound is also not achievable in partially observable settings where agents are not able to learn mappings from observations to underlying states.
2. We introduce a new approach to maximizing the mutual information between skills and observations that uses a tighter variational lower bound relative to prior work and a new actor-critic architecture.
   **Context:** None
3. We provide empirical evidence that our empowerment objective can be used to jointly learn (i) representations suitable for reinforcement learning and (ii) large sets of skills that can be executed from the learned representations.
   **Context:** None

# Empirical Bound Information-Directed Sampling for Norm-Agnostic Bandits

**Piotr M. Suder, Eric Laber**

**Keywords:** bandit algorithms, information-directed sampling, parameter bounds, heteroskedastic noise

## Summary

Information-directed sampling (IDS) is a powerful framework for solving bandit problems which has shown strong results in both Bayesian and frequentist settings. However, frequentist IDS, like many other bandit algorithms, requires that one have prior knowledge of a (relatively) tight upper bound on the norm of the true parameter vector governing the reward model in order to achieve good performance. Unfortunately, this requirement is rarely satisfied in practice. As we demonstrate, using a poorly calibrated bound can lead to significant regret accumulation. To address this issue, we introduce a novel frequentist IDS algorithm that iteratively refines a high-probability upper bound on the true parameter norm using accumulating data. We focus on the linear bandit setting with heteroskedastic subgaussian noise. Our method leverages a mixture of relevant information gain criteria to balance exploration aimed at tightening the estimated parameter norm bound and directly searching for the optimal action. We establish regret bounds for our algorithm that do not depend on an initially assumed parameter norm bound and demonstrate that our method outperforms state-of-the-art IDS and UCB algorithms.

## Contribution(s)

1. This paper introduces a novel frequentist information-directed sampling (IDS) algorithm that does not require prior knowledge of a tight upper bound of the true parameter norm to achieve good performance. Our method uses accumulating data to generate a sequence of high-probability upper bounds on the parameter norm and accounts for potential heteroskedasticity of the rewards.
   **Context:** The performance of many frequentist bandit algorithms, including various IDS (Kirschner & Krause, 2018; Kirschner et al., 2021) and UCB methods (Auer, 2002; Abbasi-Yadkori et al., 2011), relies heavily on a (at least relatively) tight upper bound on the true parameter norm being available to the algorithm. This is almost never the case in practice which can lead to significant regret accumulation. Recently, some norm-agnostic bandit algorithms have been proposed to address this issue (Gales et al., 2022), however, they do not account for potential heteroskedasticity of the rewards.
2. We introduce a new composite information gain criterion that balances improving the requisite upper bound on the parameter norm and direct search for the optimal action.
   **Context:** To the best of our knowledge, no other IDS algorithm uses a mixture of information gain criteria to balance acquiring information about different aspects of the environment's dynamics. We are also not aware of any existing method that uses an information gain criterion aimed at improving the upper bound on the parameter norm.
3. We establish anytime sublinear regret bounds for our algorithm which eventually do not depend on the initially assumed parameter norm bound.
   **Context:** Previously proposed norm-agnostic bandits (Gales et al., 2022) rely on an initial burn-in during which regret accumulation is not controlled, e.g., it need not be sublinear.

# Thompson Sampling for Constrained Bandits

**Rohan Deb, Mohammad Ghavamzadeh, Arindam Banerjee**

**Keywords:** Bandits with Knapsacks, Thompson Sampling, Conservative Bandits.

## Summary

Contextual bandits model sequential decision-making where an agent balances exploration and exploitation to maximize long term cumulative rewards. Many real-world applications, such as online advertising and inventory pricing, impose additional resource constraints while in high-stakes settings like healthcare and finance, early-stage exploration can pose significant risks. The Contextual Bandits with Knapsacks (CBwK) framework extends contextual bandits to incorporate resource constraints while the Contextual Conservative Bandit (CCB) framework ensures that performance remains above $(1+\alpha)$ times the performance of a predefined safe baseline. Although Upper Confidence Bound (UCB) based methods exist for both setups, a Thompson Sampling (TS) based approach has not been explored. This gap in the literature motivates the need to study TS for constrained settings, further reinforced by the fact that Thompson sampling often demonstrates superior empirical performance in the unconstrained setting. In this work we consider linear CBwK and CCB setups and design Thompson sampling algorithms `LinCBwK-TS` and `LinCCB-TS` respectively. We provide a $\tilde{\mathcal{O}}\big((\frac{\text{OPT}}{B}+1)m\sqrt{T}\big)$ regret for `LinCBwK-TS` where OPT is the optimal value and $B$ is the total budget. Further, we show that `LinCCB-TS` has a regret bounded by $\tilde{\mathcal{O}}\big(\sqrt{T}\min\{m^{3/2}, m\sqrt{\log K}\} + \Delta_h/\alpha r_l(\Delta_l + \alpha r_l)\big)$ and maintains the performance guarantee with high probability where $\Delta_h$ and $\Delta_l$ are the upper and lower bounds on the baseline gap and $r_l$ is a lower bound on baseline reward.

## Contribution(s)

1. We provide a Thompson Sampling Algorithm for Linear Contextual Bandits with Knapsacks and prove a high probability regret bound.
   **Context:** Previous work looked at an Upper Confidence Bound (UCB) approach.
2. We provide a Thompson Sampling Algorithm for Linear Contextual Conservative Bandits and prove a high probability regret bound along with showing that it satisfies a performance constraint.
   **Context:** Previous work looked at an Upper Confidence Bound (UCB) approach.

# AI in a vat: Fundamental limits of efficient world modelling for agent sandboxing and interpretability

**Fernando E. Rosas, Alexander Boyd, Manuel Baltieri**

**Keywords:** World models, agent sandboxing, POMDPs, AI interpretability, AI safety

## Summary

While traditionally conceived as tools to improve the task performance of model-based reinforcement learning agents, recent work has proposed *world models* as a way to build controlled virtual environments where AI agents can be thoroughly evaluated before deployment. The efficacy of these approaches, however, critically relies on the ability of world models to accurately represent real environments, which can result on high computational costs that may substantially restrict testing capabilities. Drawing inspiration from the 'brain in a vat' thought experiment, here we investigate methods to simplify world models that remain agnostic to the agent under evaluation. Our results reveal a fundamental trade-off inherent to the construction of world models related to their efficiency and interpretability. Building on this trade-off, we develop approaches that either minimise memory usage, establish the limits on what is learnable, or enable retrodictive analyses tracking the causes of undesirable outcomes. Overall, these results sheds light on the fundamental constraints that shape the design space of world modelling for agent sandboxing and interpretability.

## Contribution(s)

1. This paper conceptualises and formalises a novel problem: building efficient world models for an operator to sandbox, evaluate, and interpret AI agents before deployment.
   **Context:** Prior work (e.g. (Ha & Schmidhuber, 2018; Hafner et al., 2020)) focuses on world models from the perspective of the agent using for boosting performance, and has not considered this safety-inspired perspective.
2. We introduce generalised transducers based on quasi-probabilities, leading to a more efficient approach to compress world models at the expense of their interpretability.
   **Context:** Generalised transducers are an extension of generalised hidden Markov models, which have been thoroughly studied in previous works (Upper, 1997; Vidyasagar, 2011).
3. We provide a unifying framework to investigate and reason about world models of beliefs, and show that all models that can be calculated by an agent in real time can be bisimulated into a canonical world model known as $\epsilon$-*transducer*.
   **Context:** The minimality of the $\epsilon$-transducer among prescient rival partitions was proven in (Barnett & Crutchfield, 2015), without investigating links with bisimulation or other concepts from reinforcement learning. Relationships between bisimulation and other computational mechanics constructions were investigated by Zhang et al. (2019).
4. We introduce the notion of *reverse* interpretability, which is related to retrodictive analyses that can identify the roots of undesirable outcomes.
   **Context:** Standard interpretability approaches assess agents with respect to their capabilities to predict and plan with respect to future events (Nanda et al., 2023; Gurnee & Tegmark, 2023; Shai et al., 2025).
5. We introduce the notion of reversible transducer, and identify necessary and sufficient conditions for its construction. We also introduce and explore the notion of retrodictive beliefs.
   **Context:** Retrodictive and reversible hidden Markov models have been investigated by Ellison et al. (2009; 2011).

# Burning RED: Unlocking Subtask-Driven Reinforcement Learning and Risk-Awareness in Average-Reward Markov Decision Processes

**Juan Sebastian Rojas, Chi-Guhn Lee**

**Keywords:** Average-Reward Reinforcement Learning, Risk-Sensitive Decision-Making, CVaR

## Summary

Average-reward Markov decision processes (MDPs) provide a foundational framework for sequential decision-making under uncertainty. However, average-reward MDPs have remained largely unexplored in reinforcement learning (RL) settings, with the majority of RL-based efforts having been allocated to discounted MDPs. In this work, we study a unique structural property of average-reward MDPs and utilize it to introduce *Reward-Extended Differential* (or *RED*) reinforcement learning: a novel RL framework that can be used to effectively and efficiently solve various learning objectives, or *subtasks*, simultaneously in the average-reward setting. We introduce a family of RED learning algorithms for prediction and control, including proven-convergent algorithms for the tabular case. We then showcase the power of these algorithms by demonstrating how they can be used to learn a policy that optimizes, for the first time, the well-known conditional value-at-risk (CVaR) risk measure in a fully-online manner, *without* the use of an explicit bi-level optimization scheme or an augmented state-space.

## Contribution(s)

1. We provide a general-purpose framework and a corresponding set of prediction/control algorithms for solving an arbitrary number of learning objectives, or *subtasks*, simultaneously in the average-reward setting with only a TD error-based update, including proven-convergent algorithms for the tabular case.
   **Context:** Our work builds on (and can be viewed as a generalization of) Wan et al. (2021), which proposed proven-convergent average-reward RL algorithms that are able to learn and/or optimize the value function and average-reward simultaneously using only the TD error. In particular, the focus in Wan et al. (2021) was on proving the convergence of such algorithms, without exploring the underlying structural properties of the average-reward MDP that made such a process possible to begin with. In this work, we formalize these underlying properties, and utilize them to show that if one modifies, or *extends*, the reward from the MDP with various learning objectives, then these objectives, or *subtasks*, can be solved simultaneously using a modified, or *reward-extended*, version of the TD error.
2. We utilize the framework described in Contribution 1 to derive the first family of RL algorithms that can optimize the well-known conditional value-at-risk (CVaR) risk measure in a fully-online manner *without* the use of an explicit bi-level optimization scheme or an augmented state-space. We perform an empirical evaluation on two toy experiments, thereby illustrating the properties and effectiveness of the algorithms, while also noting that a more comprehensive empirical study is needed to fully gauge their practical implications.
   **Context:** Several prior works have investigated CVaR optimization in the discounted setting (e.g. Bäuerle and Ott (2011) and Chow et al. (2015)). However, no prior work has developed an algorithm for CVaR optimization that does not require either an augmented state-space or an explicit bi-level optimization, which can, for example, involve solving multiple MDPs. In the average-reward setting, Xia et al. (2023) proposed a set of algorithms for optimizing the CVaR risk measure, however their methods require the use of an augmented state-space and a sensitivity-based bi-level optimization. By contrast, our work, to the best of our knowledge, is the first to optimize CVaR in an MDP-based setting without the use of an explicit bi-level optimization scheme or an augmented state-space.

# Achieving Limited Adaptivity for Multinomial Logistic Bandits

**Sukruta Prakash Midigeshi, Tanmay Goyal, Gaurav Sinha**

**Keywords:** Multinomial Logistic Bandits, Limited Adaptivity, Batched Bandits, Contextual Bandits

## Summary

Multinomial Logistic Bandits have recently attracted much attention due to their ability to model problems with multiple outcomes. In this setting, each decision is associated with many possible outcomes, modeled using a multinomial logit function. Several recent works on multinomial logistic bandits have simultaneously achieved optimal regret and computational efficiency. However, motivated by real-world challenges and practicality, there is a need to develop algorithms with limited adaptivity, wherein we are allowed only $M$ policy updates. To address these challenges, we present two algorithms, `B-MNL-CB` and `RS-MNL`, that operate in the batched and rarely-switching paradigms, respectively. The batched setting involves choosing the $M$ policy update rounds at the start of the algorithm, while the rarely-switching setting can choose these $M$ policy update rounds in an adaptive fashion. Our first algorithm, `B-MNL-CB` extends the notion of distributional optimal designs to the multinomial setting and achieves $\tilde{O}(\sqrt{T})$ regret assuming the contexts are generated stochastically when presented with $\Omega(\log\log T)$ update rounds. Our second algorithm, `RS-MNL` works with adversarially generated contexts and can achieve $\tilde{O}(\sqrt{T})$ regret with $\tilde{O}(\log T)$ policy updates. Further, we conducted experiments that demonstrate that our algorithms (with a fixed number of policy updates) are extremely competitive (and often better) than several state-of-the-art baselines (which update their policy every round), showcasing the applicability of our algorithms in various practical scenarios.

## Contribution(s)

1. We present an algorithm, `B-MNL-CB`, that achieves an optimal $\tilde{O}(\sqrt{T})$ regret with $\Omega(\log\log T)$ batches in the batched setting. Moreover, the leading term of the regret is independent of $\kappa$, an instance-dependent non-linearity parameter.
   **Context:** In the batched setting, the rounds at which the policy is updated are fixed beforehand. Gao et al. (2019) showed that having $\Omega(\log\log T)$ batches is necessary to achieve the optimal minimax regret. Our algorithm, `B-MNL-CB`, combines the idea of distributional optimal designs (introduced in Ruan et al. (2021)) with the idea of suitable scalings for arms (introduced in Sawarni et al. (2024)) to the multinomial logistic setting. This requires a natural extension of distributional optimal designs to this setting. Achieving a $\kappa-$ independent regret is important because Amani & Thrampoulidis (2021) showed that $\kappa$ scales exponentially in several instance parameters and hence, can increase the regret significantly.
2. We present a rarely-switching algorithm `RS-MNL` that achieves an optimal $\tilde{O}(\sqrt{T})$ regret (with a $\kappa-$free leading term) requiring $O(\log T)$ switches (policy updates).
   **Context:** In the rarely-switching setting, the switching rounds (policy-update rounds) are adaptively chosen during the course of the algorithm. The need for the update is decided based on a switching criterion similar to the one in Abbasi-Yadkori et al. (2011). While the algorithm bears similarities to the rarely-switching algorithm presented in Sawarni et al. (2024), an alternate regret decomposition method allows us to get rid of the warm-up criterion, which helps reduce the number of switches from $O(\log^2 T)$ to $O(\log T)$. Further, we also get rid of the *Successive Eliminations* in Sawarni et al. (2024) that determine the arm to be played, and replace it with the simpler UCB-maximization rule of Abbasi-Yadkori et al. (2011), resulting in a more efficient runtime for the algorithm.

# Which Experiences Are Influential for RL Agents? Efficiently Estimating The Influence of Experiences

**Takuya Hiraoka, Guanquan Wang, Takashi Onishi, Yoshimasa Tsuruoka**

**Keywords:** reinforcement learning, data influence estimation

## Summary

In reinforcement learning (RL) with experience replay, experiences stored in a replay buffer influence the RL agent's performance. Information about how these experiences influence the agent's performance is valuable for various purposes, such as identifying experiences that negatively influence underperforming agents. One method for estimating the influence of experiences is the leave-one-out (LOO) method. However, this method is usually computationally prohibitive. In this paper, we present Policy Iteration with Turn-over Dropout (PIToD), which efficiently estimates the influence of experiences. We evaluate how correctly PIToD estimates the influence of experiences and its efficiency compared to LOO. We then apply PIToD to amend underperforming RL agents, i.e., we use PIToD to estimate negatively influential experiences for the RL agents and to delete the influence of these experiences. We show that RL agents' performance is significantly improved via amendments with PIToD. Our code is available at: https://github.com/TakuyaHiraoka/Which-Experiences-Are-Influential-for-RL-Agents

## Contribution(s)

1. For the first time, we propose a method that efficiently (i) estimates the influence of individual experiences (i.e., data) on the performance (e.g., empirical returns) of an RL agent and (ii) disables that influence when necessary (Section 4).

   **Why is this contribution valuable?** In many RL settings, we must manage experiences of different quality levels. For example, in an off-policy RL setting, experiences collected from multiple policies—ranging from random to near-optimal—are used to learn policies or Q-functions. When experiences of different quality are intermixed, the ability to estimate their influence on performance and disable any harmful influences is highly beneficial for many purposes. For instance, (i) if an RL agent's performance is degraded by specific detrimental experiences, disabling their influence can help improve the agent's overall performance. (ii) In safety-critical applications (e.g., human-in-the-loop robotics or autonomous driving), this ability may ensure safety by disabling the influence of experiences that degrade safety performance before deployment. In addition, (iii) in memory-intensive scenarios like image-based RL, where each experience consumes substantial computational memory, this ability may enable efficient management of experiences by screening out less useful experiences. Finally, (iv) when refining RL task design, analyzing influential experiences may provide valuable insights for improving reward functions or state representations, leading to better task design.

   **Context:** (i) No prior work has addressed the efficient estimation and disabling of the influence of experiences in the online RL context. (ii) As a first step, this paper focuses on verifying the proposed method's effectiveness within single-task off-policy RL settings (MuJoCo and DMC) (Section 5 and 6, and Appendix H).

# Your Learned Constraint is Secretly a Backward Reachable Tube

**Mohamad Qadri**[1], **Gokul Swamy**[1], **Jonathan Francis**[1,2], **Michael Kaess**[1], **Andrea Bajcsy**[1]

**Keywords:** Constraint Inference, Learning from Demonstration, Safe Control

## Summary

Inverse Constraint Learning (ICL) is the problem of inferring constraints from safe (i.e., constraint-satisfying) demonstrations. The hope is that these inferred constraints can then be used downstream to search for safe policies for new tasks and, potentially, under different dynamics. Our paper explores the question of what mathematical entity ICL recovers. Somewhat surprisingly, we show that both in theory and in practice, ICL recovers the set of states where failure is *inevitable*, rather than the set of states where failure has *already* happened. In the language of safe control, this means we recover a *backwards reachable tube (BRT)* rather than a *failure set*. In contrast to the failure set, the BRT depends on the dynamics of the data collection system. We discuss the implications of the dynamics-conditionedness of the recovered constraint on both the sample-efficiency of policy search and the transferability of learned constraints. Our code is available in the following repository.

## Contribution(s)

1. This paper establishes a connection between Inverse Constraint Learning and Hamilton-Jacobi (HJ) Reachability from safe control theory, providing a new theoretical perspective on learning constraints from demonstrations.
   **Context:** None
2. We prove theoretically and verify experimentally that the mathematical set encoded by the learned constraint is a dynamics-dependent Backward Reachable Tube (BRT) and not the dynamics independent Failure Set.
   **Context:** Prior works implicitly assume that the constraint learned via ICL is dynamics independent. In this paper we show that the constraint will actually depend on the dynamics of the expert demonstrators.
3. We discuss the implication of this observation in terms of the sample-efficiency of policy search and transferability of the learned constraint.
   **Context:** None

# Improved Regret Bound for Safe Reinforcement Learning via Tighter Cost Pessimism and Reward Optimism

**Kihyun Yu, Duksang Lee, William Overman, Dabeen Lee**

**Keywords:** Safe Reinforcement Learning, Constrained MDPs, Regret Analysis.

## Summary

This paper studies the safe reinforcement learning problem formulated as an episodic finite-horizon tabular constrained Markov decision process with an unknown transition kernel and stochastic reward and cost functions. We propose a model-based algorithm based on novel cost and reward function estimators that provide tighter cost pessimism and reward optimism. While guaranteeing no constraint violation in every episode, our algorithm achieves a regret upper bound of $\widetilde{\mathcal{O}}((\bar{C}-\bar{C}_b)^{-1}H^{2.5}S\sqrt{AK})$ where $\bar{C}$ is the cost budget for an episode, $\bar{C}_b$ is the expected cost under a safe baseline policy over an episode, $H$ is the horizon, and $S$, $A$ and $K$ are the number of states, actions, and episodes, respectively. This improves upon the best-known regret upper bound, and when $\bar{C}-\bar{C}_b=\Omega(H)$, the gap from the regret lower bound of $\Omega(H^{1.5}\sqrt{SAK})$ is $\widetilde{\mathcal{O}}(\sqrt{S})$. We deduce our cost and reward function estimators via a Bellman-type law of total variance to obtain tight bounds on the expected sum of the variances of value function estimates. This leads to a tighter dependence on the horizon in the function estimators. We also present numerical results to demonstrate the computational effectiveness of our proposed framework.

## Contribution(s)

1. This paper presents an algorithm for episodic finite-horizon tabular constrained Markov decision processes with an improved regret upper bound of $\widetilde{\mathcal{O}}((\bar{C}-\bar{C}_b)^{-1}H^{2.5}S\sqrt{AK})$, ensuring zero constraint violation over all episodes.
   **Context:** The best-known regret upper bound is $\widetilde{\mathcal{O}}((\bar{C}-\bar{C}_b)^{-1}H^{3}S\sqrt{AK})$ due to Bura et al. (2022), and our result improves it by a factor of $\widetilde{\mathcal{O}}(\sqrt{H})$. The zero constraint violation setting means that there is no episode in which the constraint is violated. Additionally, a safe baseline policy is assumed to be known as in Liu et al. (2021); Bura et al. (2022).
2. When $\bar{C}-\bar{C}_b=\Omega(H)$, our algorithm is the first algorithm that nearly matches the lower bound of $\Omega(H^{1.5}\sqrt{SAK})$ in terms of $H$ in the zero constraint violation setting.
   **Context:** The lower bound is originally derived for the unconstrained case (Jin et al., 2020; Domingues et al., 2021), and it also works for the constrained case as we can take trivial cost functions.
3. The key is to control the error of estimating the unknown transition kernel over each episode. In particular, we provide a tighter bound on the estimation error for each episode, based on a Bellman-type law of total variance.
   **Context:** Our Bellman-type law of total variance technique refines the analysis of Bura et al. (2022), resulting in a tighter bound expressed as a function of the estimated transition kernel. The technique is inspired by Chen & Luo (2021), while they gave only a cumulative error bound across all episodes, and at the same time, the bound is expressed as a function of the true transition kernel which is unknown to the algorithm.

# Offline vs. Online Learning in Model-based RL: Lessons for Data Collection Strategies

**Jiaqi Chen , Ji Shi , Cansu Sancaktar , Jonas Frey , Georg Martius**

**Keywords:** Model-based RL, Online Learning, Offline Learning, Active Learning, Exploration

## Summary

Data collection is crucial for learning robust world models in model-based reinforcement learning. The most prevalent strategies are to actively collect trajectories by interacting with the environment during online training or training on offline datasets. At first glance, the nature of learning task-agnostic environment dynamics makes world models a good candidate for effective offline training. However, the effects of online vs. offline data on world models and thus on the resulting task performance have not been thoroughly studied in the literature. In this work, we investigate both paradigms in model-based settings, conducting experiments on 31 different environments. First, we showcase that online agents outperform their offline counterparts. We identify a key challenge behind performance degradation of offline agents: encountering Out-Of-Distribution (OOD) states at test time. This issue arises because, without the self-correction mechanism in online agents, offline datasets with limited state space coverage induce a mismatch between the agent's imagination and real rollouts, compromising policy training. We demonstrate that this issue can be mitigated by allowing for additional online interactions in a fixed or adaptive schedule, restoring the performance of online training with limited interaction data. We also showcase that incorporating exploration data helps mitigate the performance degradation of offline agents. Based on our insights, we recommend adding exploration data when collecting large datasets, as current efforts predominantly focus on expert data alone.

## Contribution(s)

1. We provide an in-depth analysis of performance degradation in offline model-based agents with practical considerations. We highlight the coupling of model and policy learning as a primary contributing factor beyond the pure OOD challenge.
   **Context:** In model-free RL, the performance degradation is often linked to limited coverage of offline datasets, which leads to inaccurate value estimates and poor extrapolation of the learned policy (Ostrovski et al., 2021; Yue et al., 2023; 2022). Similar issues plague offline model-based RL (He, 2023; Kidambi et al., 2020; Chen et al., 2023; Yu et al., 2020; Cang et al., 2021). However, an alternative perspective remains overlooked: the influence of data quality and online interaction ratios on the robustness and generalization of world models.
2. We demonstrate that incorporating exploration data with a mixed reward improves the state-space coverage in offline training. This provides insights in how to create the offline dataset such that the performance degradation can be mitigated and competitive task performance can be maintained.
   **Context:** Existing methods primarily focus on constraining the agent within in-distribution regions for the task (Kidambi et al., 2020; Yu et al., 2020; 2021; Wang et al., 2024; Matsushima et al., 2021) but do not explicitly assess which data collection strategies best support offline training.
3. We propose using the world model loss as a metric to measure the novelty of regions explored by the current policy. It serves as an indicator for when minimal online interactions can help offline agents to efficiently improve performance.
   **Context:** The approach of self-generated data is mostly investigated in the context of model-free RL (Ostrovski et al., 2021; Lee et al., 2021).

# Uncertainty Prioritized Experience Replay

**Rodrigo Carrasco-Davis, Sebastian Lee, Claudia Clopath, Will Dabney**

**Keywords:** Experience Replay, Uncertainty Estimation, Information Gain

## Summary

Prioritized experience replay, which improves sample efficiency by selecting relevant transitions to update parameter estimates, is a crucial component of contemporary value-based deep reinforcement learning models. Typically, transitions are prioritized based on their temporal difference error. However, this approach is prone to favoring noisy transitions, even when the value estimation closely approximates the target mean. This phenomenon resembles the *noisy TV* problem postulated in the exploration literature, in which exploration-guided agents get stuck by mistaking noise for novelty. To mitigate the disruptive effects of noise in value estimation, we propose using epistemic uncertainty to guide the prioritization of transitions from the replay buffer. Epistemic uncertainty quantifies the uncertainty that can be reduced by learning, hence reducing transitions sampled from the buffer generated by unpredictable random processes. We first illustrate the benefits of epistemic uncertainty prioritized replay in two tabular toy models: a simple multi-arm bandit task, and a noisy gridworld. Subsequently, we evaluate our prioritization scheme on the Atari suite, outperforming quantile regression deep Q-learning benchmarks; thus forging a path for the use of epistemic uncertainty prioritized replay in reinforcement learning agents.

## Contribution(s)

1. We introduce a new decomposition of uncertainties in reinforcement learning extending previous formulations of epistemic and aleatoric uncertainty estimators (Clements et al., 2020) to include a distance-to-target term. This decomposition better accounts for bias-variance trade-offs in the underlying estimator.
   **Context:** While Clements et al. (2020) start by defining total uncertainty estimator as the variance over distributional and ensemble dimensions of the value estimate, we start instead from the average square error to the target over distributional and ensemble dimensions. Under the definitions given by Lahlou et al. (2022) in their Direct Epistemic Uncertainty Prediction (DEUP) framework, this yields a modified epistemic uncertainty estimator that we term the *target epistemic uncertainty*.
2. We propose using these measures of epistemic and aleatoric uncertainty in an *information gain* criterion to prioritize experience replay in reinforcement learning. We call this prioritization scheme Uncertainty Prioritized Experience Replay (UPER).
   **Context:** The de facto method for prioritizing replay in reinforcement learning has been the absolute value of the temporal difference error since its introduction by Schaul et al. (2016). However we argue that this can lead to sub-optimal behavior in noisy environments. We go on to derive the information gain prioritization criterion from principled treatment of a toy Bayesian problem.
3. We demonstrate the effectiveness of this prioritization scheme in two toy models (a bandit and gridworld), as well as in a deep learning model on the Atari test suite. In the latter we use an ensemble of distributional QR agents (Dabney et al., 2017) to estimate the relevant uncertainty quantities.
   **Context:** We provide a series of ablation studies in Atari that isolate the effect of the prioritization variable (from architectural changes such as adding an ensemble), showing that UPER could be a promising alternative to PER and other uncertainty measures like plain ensemble disagreement.

# RL$^3$: Boosting Meta Reinforcement Learning via RL inside RL$^2$

**Abhinav Bhatia, Samer B. Nashed, Shlomo Zilberstein**

**Keywords:** Meta-reinforcement learning

## Summary

Meta reinforcement learning (Meta-RL) methods such as RL$^2$ have emerged as promising approaches for learning data-efficient RL algorithms tailored to a given task distribution. However, they show poor asymptotic performance and struggle with out-of-distribution tasks because they rely on sequence models, such as recurrent neural networks or transformers, to process experiences rather than summarize them using general-purpose RL components such as value functions. In contrast, traditional RL algorithms are data-inefficient as they do not use domain knowledge, but do converge to an optimal policy in the limit. We investigate the hypothesis that incorporating action-values, learned per task via traditional RL, in the inputs to Meta-RL systems that solve the meta-level decision process via an 'outer-loop' deep RL algorithm and a sequence model (e.g. recurrent network, transformer) has a positive effect on the above shortcomings. Using an example implementation, called RL$^3$, we demonstrate that this strategy earns greater cumulative reward in the long term compared to RL$^2$ while drastically reducing meta-training time and generalizing better to out-of-distribution tasks. Experiments are conducted on both custom and benchmark discrete domains from the Meta-RL literature that exhibit a range of short-term, long-term, and complex dependencies.

## Contribution(s)

1. A thorough investigation of the hypothesis that augmenting Meta-RL inputs with task specific $Q$-value estimates improves performance across several metrics, which to some may be surprising as this information is already latent within the original input sequence.
   **Context:** Although some results we present are strong, *this paper is not attempting to present a state-of-the-art Meta-RL system across the board*. We are interested in determining the effects of augmenting the typical Meta-RL inputs of sequence models solving the 'outer-loop' using deep RL algorithms with task specific $Q$-value estimates, without using other privileged information or extra resources which in practice often increase performance significantly and are in theory *compatible* with this work.
2. This paper presents thorough theoretical, empirical, and logical arguments for the effectiveness of $Q$-estimate state-augmentation; significant attention is devoted to understanding, from different perspectives, why this method achieves the results we see empirically.
   **Context:** Previous papers have at times highlighted the importance of object-level value estimation for successful Meta-RL, though never in the form of an algorithm such as RL$^3$. This paper presents a unique method, set of experiments, and additional analysis complementing existing research in our effort to better understand the capabilities and properties of Meta-RL systems that learn to solve the meta-level decision process via deep RL algorithms with sequence models (e.g. recurrent networks, transformers).

# Pareto Optimal Learning from Preferences with Hidden Context

**Ryan Bahlous-Boldi, Li Ding, Lee Spector, Scott Niekum**

**Keywords:** Preference Learning, Pareto-optimality, Lexicase Selection, Hidden Context

## Summary

Ensuring AI models align with human values is essential for their safety and functionality. Reinforcement learning from human feedback (RLHF) leverages human preferences to achieve this alignment. However, when preferences are sourced from diverse populations, point estimates of reward can result in suboptimal performance or be unfair to specific groups. We propose Pareto Optimal Preference Learning (POPL), which enables pluralistic alignment by framing discrepant group preferences as objectives with potential trade-offs, aiming for policies that are Pareto-optimal on the preference dataset. POPL utilizes lexicase selection, an iterative process that selects diverse and Pareto-optimal solutions. Our theoretical and empirical evaluations demonstrate that POPL surpasses baseline methods in learning sets of reward functions and policies, effectively catering to distinct groups without access to group numbers or membership labels. We verify the performance of POPL on a stateless preference learning setting, a Minigrid RL domain, Metaworld robotics benchmarks, as well as large language model (LLM) fine-tuning. We illustrate that POPL can also serve as a foundation for techniques optimizing specific notions of group fairness, ensuring safe and equitable AI model alignment.

## Contribution(s)

1. We extend the problem of Reinforcement Learning from Human Feedback with Hidden Context (RLHF-HC) introduced by Siththaranjan et al. (2023), addressing critical limitations in preference learning for sequential, time-based domains, as opposed to contextual bandits.
   **Context:** Siththaranjan et al. (2023) assumes a contextual bandit setting, where hidden context exists independently across states. For use in sequential settings, we argue that preference learning frameworks must pay attention to persistent annotator identity.
2. We adapt lexicase selection to preference learning, enabling an iterative process to filter candidate models based on diverse subsets of human preferences.
   **Context:** Lexicase selection has been used in a variety of other domains (Spector et al., 2024); here, it is adapted to handle conflicting human preferences in sequential RL settings.
3. We provide theoretical justification showing that, under noiseless conditions, optimal reward functions and policies for hidden context groups are inherently Pareto-Optimal with respect to the the entire set of preferences.
   **Context:** This result grounds the method in robust multi-objective optimization principles, offering clear theoretical support for POPL, while acknowledging that real-world settings will need to manage additional complexities such as noise and choice of regularization.
4. We empirically demonstrate that searching for Pareto-optimal reward functions and policies recovers those that align with the values of specific groups of humans.
   **Context:** We show this by creating situations where hidden context will be present in a variety of tasks, including Minigrid (Chevalier-Boisvert et al., 2023), Metaworld (Yu et al., 2019) and LLM jailbreaking detection based on RLHF-HH (Bai et al., 2022; Wei et al., 2024), and show that our reward or policy inference set contains personalized models for our chosen groups.

# WOFOSTGym: A Crop Simulator for Learning Annual and Perennial Crop Management Strategies

**William Solow, Sandhya Saisubramanian, Alan Fern**

**Keywords:** Crop Simulator, Reinforcement Learning for Agriculture

## Summary

We introduce WOFOSTGym, a novel crop simulation environment designed to train reinforcement learning (RL) agents to optimize agromanagement decisions for annual and perennial crops in multi-farm settings. Effective crop management requires optimizing yield and economic returns while minimizing environmental impact, which is a complex sequential decision-making problem well-suited for RL. However, the lack of simulators for perennial crops in multi-farm contexts has hindered RL applications in this domain. Existing crop simulators also do not support multiple annual crops. WOFOSTGym addresses the shortcomings of available crop simulators by supporting 23 annual crops and two perennial crops, enabling RL agents to learn diverse agromanagement strategies in multi-year, multi-crop, and multi-farm settings. Our simulator offers a suite of challenging tasks for learning under partial observability, non-Markovian dynamics, and delayed feedback. Our extensive experiments across a wide variety of crops in single and multi-farm settings, including the constrained optimization tasks that arise in agriculture, demonstrate the learning capabilities and challenges of RL and imitation learning agents. The experiments highlight WOFOSTGym's potential for advancing core RL research and RL-driven decision support in agriculture.

## Contribution(s)

1. We introduce WOFOSTGym, an RL simulator built on the WOFOST crop growth model, designed for developing agromanagement policies across multiple annual and multi-season perennial crops, advancing AI-driven decision support in agriculture.
   **Context:** Existing crop simulators do not support perennial crops or multiple annual crops. WOFOSTGym addresses this gap, enabling users without agricultural expertise to create experiments with multiple farms and multiple crops, across a range of tasks with varying observability to reflect real world sensing challenges.
2. We modify the WOFOST crop growth model (CGM) to simulate the growth of perennial crops across multiple growing seasons, and update WOFOST nutrient modules to be able to investigate the impact of agromanagement decisions on the surrounding environment.
   **Context:** Bai et al. (2019) used the WOFOST crop growth model (CGM) to model the growth of the perennial jujube tree across multiple seasons. Inspired by their work, we modified the WOFOST CGM to support perennial growth within WOFOSTGym, and to model continuous multi-year growth with the addition of a dormancy phase.
3. We apply Bayesian Optimization to calibrate the parameters of the WOFOST CGM to increase model fidelity and compare our results with those of an existing work that collected phenology data for 10 grape cultivars.
   **Context:** High-fidelity CGMs are essential for sim-to-real transfer in open-field agriculture, but parameter calibration is challenging and time-consuming. Traditional agronomic methods rely on linear regression or Monte Carlo sampling. In contrast, our Bayesian Optimization approach provides a more efficient, principled search of the CGM parameter space, achieving comparable or superior results with fewer computations and limited field data.

# When and Why Hyperbolic Discounting Matters for Reinforcement Learning Interventions

**Ian M. Moore, Eura Nofshin, Siddharth Swaroop, Susan Murphy, Finale Doshi-Velez, Weiwei Pan**

**Keywords:** Hyperbolic discounting, Human-AI interaction, Agent-based modeling of humans

## Summary

In settings where an AI agent sends interventions to nudge a human agent toward a goal, the AI's ability to quickly learn a high-quality policy depends on how well it models the human. Despite behavioral evidence that humans hyperbolically discount future rewards, we continue to model human agents as Markov Decision Processes (MDPs) with exponential discounting because of its mathematical properties. In this work, we derive an exponential discount factor that will never miss a necessary intervention–and minimizes unnecessary extra interventions–even when the real human is hyperbolic. In addition, we demonstrate that when the dynamics are unknown, using our exponential alternative outperforms correctly modeling the human, even when the human's true hyperbolic discount is known.

## Contribution(s)

1. Using theory, we connect model misspecification of a hyperbolic human agent as an exponential one to errors in the downstream AI intervention policy.
   **Context:** Prior work in human-AI settings has not studied how misspecifications of the human agent's discount affect AI policies. Our analysis is in the context of absorbing state MDPS (discrete state / action spaces with absorbing reward states) and on interventions of the human agent's discount factor. We make simplifying assumptions– about the stochasticity of the transitions, intermediate rewards, and noise in the human policy– which *we relax* in our empirical experiments. All humans in our experiments are simulated agents modeled using a Markov Decision Process (MDP).
2. We prove that the exponential mean hazard rate, $\gamma_{\text{mhr}}$, guarantees no false negatives in the AI policy. However, it does not minimize AI false positives.
   **Context:** The AI policy is the optimal policy for an MDP in which the actions are interventions, delivered by an artificial agent, on a human agent's MDP parameters. The mean hazard rate (MHR) is an established method for approximating hyperbolic human agents as exponential ones (Rambaud & Torrecillas, 2005; Sozou, 1998; 2009). Previously, there were no formal guarantees on how the MHR affects error when used to model human agents in a human-AI setting. The same context from contribution 1 (about absorbing-state MDPs, theoretical assumptions), apply.
3. We derive a fixed exponential discount rate, $\gamma_{\text{safe}}$, for approximating hyperbolic agents.
   **Context:** Our theoretical justification relies on the same assumptions as contribution 1. However, $\gamma_{\text{safe}}$ is as broad as $\gamma_{\text{mhr}}$ and is applicable to settings beyond the ones considered in this paper.
4. In empirical experiments (on small tabular MDPs), we demonstrate that (biased) exponential approximations using a fixed discount parameter outperform several different (unbiased) methods of approximating the hyperbolic discount when the transitions are learned online.
   **Context:** Prior work had not considered how the choice of discount model for the human agent affects the AI policy. We found that the hyperbolic approximations are unexpectedly sensitive to online learning. Our experiments are in small, tabular MDP settings.
5. Empirically, we characterize situations where a fixed exponential discount model with $\gamma_{\text{safe}}$ is preferable to a fixed one with $\gamma_{\text{mhr}}$; we do the same for $\gamma_{\text{safe}}$ vs. updating $\gamma$ online.
   **Context:** None.

# Online Intrinsic Rewards for Decision Making Agents from Large Language Model Feedback

**Qinqing Zheng, Mikael Henaff, Amy Zhang, Aditya Grover, Brandon Amos**

**Keywords:** intrinsic motivation, exploration, sparse rewards, LLMs

## Summary

Automatically synthesizing dense rewards from natural language descriptions is a promising paradigm in reinforcement learning (RL), with applications to sparse reward problems, open-ended exploration, and hierarchical skill design. Recent works have made promising steps by exploiting the prior knowledge of large language models (LLMs). However, these approaches suffer from important limitations: they are either not scalable to problems requiring billions of environment samples, due to requiring LLM annotations for each observation, or they require a diverse offline dataset, which may not exist or be impossible to collect. In this work, we address these limitations through a combination of algorithmic and systems-level contributions. We propose `ONI`, a distributed architecture that simultaneously learns an RL policy and an intrinsic reward function using LLM feedback. Our approach annotates the agent's collected experience via an asynchronous LLM server, which is then distilled into an intrinsic reward model. We explore a range of algorithmic choices for reward modeling with varying complexity, including hashing, classification, and ranking models. Our approach achieves state-of-the-art performance across a range of challenging, sparse reward tasks from the NetHack Learning Environment in a simple unified process, solely using the agent's gathered experience, without requiring external datasets.

## Contribution(s)

1. This paper presents a method, system design and open-source codebase for learning intrinsic rewards from LLM feedback in an online manner, which scales to high-throughput settings.
   **Context:** Several methods for producing intrinsic rewards from LLM feedback have been proposed in prior work. However, they either do not scale to high-throughput settings, or they require a large and diverse offline dataset. Our method works in high-throughput settings without requiring offline data.

# Reinforcement Learning from Human Feedback with High-Confidence Safety Constraints

**Yaswanth Chittepu , Blossom Metevier , Will Schwarzer , Austin Hoag , Scott Niekum , Philip S. Thomas**

**Keywords:** Language model alignment, Reinforcement learning from human feedback (RLHF), Safe reinforcement learning, AI safety

## Summary

Existing approaches to language model alignment often treat safety as a tradeoff against helpfulness, which can lead to unacceptable responses in sensitive domains. To ensure reliable performance in such settings, we propose High-Confidence Safe Reinforcement Learning from Human Feedback (HC-RLHF), a method that provides high-confidence safety guarantees while maximizing helpfulness. Similar to previous methods, HC-RLHF explicitly decouples human preferences regarding helpfulness and harmlessness (safety) and trains separate reward and cost models, respectively. It then employs a two-step process to find safe solutions. In the first step, it optimizes the reward function while ensuring that a specific upper-confidence bound on the cost constraint is satisfied. In the second step, the trained model undergoes a safety test to verify that its performance satisfies a separate upper-confidence bound on the cost constraint.

## Contribution(s)

1. We introduce HC-RLHF, the first Seldonian algorithm (Thomas et al., 2019) with applications to RLHF. With high probability, HC-RLHF can find solutions that satisfy the safety constraint introduced by Safe RLHF (Dai et al., 2023).
   **Context:** HC-RLHF builds on two works: Safe RLHF (Dai et al., 2023) and the Seldonian framework (Thomas et al., 2019). Like previous Seldonian algorithms, HC-RLHF follows a two-step process, consisting of an optimization step followed by a safety step. The optimization step in HC-RLHF is designed similarly to Safe RLHF in that it separates human preference data into two distinct objectives: helpfulness and harmlessness. The harmlessness objective is similarly treated as a constraint while optimizing for helpfulness. However, we introduce an important modification to this constraint: it is redefined to increase the likelihood that the learned model passes the safety test.
2. We provide a theoretical analysis of HC-RLHF, including a proof that it will not return an unsafe solution with a probability greater than a user-specified threshold.
   **Context:** This ensures that HC-RLHF is indeed a Seldonian algorithm (Thomas et al., 2019).
3. Empirically, we apply HC-RLHF to align three different language models (Qwen2-1.5B, Qwen2.5-3B, and LLaMa-3.2-3B) with human preferences. Our results demonstrate that HC-RLHF produces safe models with high probability while also improving helpfulness and harmlessness compared to previous methods.
   **Context:** We use the dataset used by Dai et al. (2023), and compare the helpfulness and harmlessness of models trained by HC-RLHF, Safe RLHF, and Supervised Fine Tuning.

# AVID: Adapting Video Diffusion Models to World Models

**Marc Rigter, Tarun Gupta, Agrin Hilmkil, Chao Ma**

**Keywords:** world models, diffusion models, model-based reinforcement learning, black-box adaptation

## Summary

Reinforcement learning (RL) is highly effective in domains that can be easily simulated. However, in problems such as robotic manipulation, accurate simulation is challenging and gathering large amounts of real-world data is impractical. A potential solution lies in leveraging widely-available unlabelled videos to train world models that simulate the consequences of actions. If the world model is accurate, it can be used to generate synthetic data to optimize decision-making via RL. Image-to-video diffusion models are already capable of generating highly realistic synthetic videos. However, these models are not action-conditioned, and the most powerful models are closed-source which means they cannot be finetuned. In this work, we propose to adapt pretrained video diffusion models to action-conditioned world models, without access to the parameters of the pretrained model. Our approach, AVID, trains an adapter on a small domain-specific dataset of action-labelled videos. AVID uses a learned mask to modify the intermediate outputs of the pretrained model and generate accurate action-conditioned videos. We evaluate AVID on video game and real-world robotics data, and show that it generally outperforms baselines for diffusion adaptation in video and image metrics. AVID demonstrates that pretrained video models have the potential to be powerful tools for generating synthetic data for RL agents. In future work, we wish to investigate how the improved data generation accuracy translates to model-based RL performance.

## Contribution(s)

1. Proposing to adapt pre-trained video diffusion models to action-conditioned world models under the assumption that we do not have access to the parameters of the pretrained model, but can access the pretrained model outputs at each diffusion step.
   **Context:** Existing work on adapting video models to action-conditioned world models finetunes the original model (Seo et al., 2022; Wu et al., 2024). In our work, we assume that the pretrained model cannot be finetuned and we only have access to its outputs at each diffusion step. The goal is to train an adapter that modifies these outputs so that the resulting model can generate accurate action-conditioned samples that are suitable for use in model-based RL.
2. AVID, a novel approach to adding conditioning to pretrained diffusion models. AVID applies a learned mask to the outputs of a pretrained model, and combines them with action-conditioned outputs generated by a domain-specific adapter.
   **Context:** AVID can be thought of as learning an adapter that "guides" the original video model towards a valid action-conditioned sample. We demonstrate that AVID performs well in terms of image and video accuracy metrics for the synthetic videos produced. However, we do not evaluate the performance of using this data to train RL agents in the model-based RL setting. We wish to do this in future work.

# Non-Stationary Latent Auto-Regressive Bandits

**Anna L. Trella, Walter Dempsey, Asim H. Gazi, Ziping Xu, Finale Doshi-Velez, Susan A. Murphy**

**Keywords:** bandit algorithms, non-stationarity

## Summary

For the non-stationary multi-armed bandit (MAB) problem, many existing methods allow a general mechanism for the non-stationarity, but rely on a budget for the non-stationarity that is sub-linear to the total number of time steps $T$. In many real-world settings, however, the mechanism for the non-stationarity can be modeled, but there is no budget for the non-stationarity. We instead consider the non-stationary bandit problem where the reward means change due to a latent, auto-regressive (AR) state. We develop Latent AR LinUCB (LARL), an online linear contextual bandit algorithm that does not rely on the non-stationary budget, but instead forms predictions of reward means by implicitly predicting the latent state. The key idea is to reduce the problem to a linear dynamical system which can be solved as a linear contextual bandit. In fact, LARL approximates a steady-state Kalman filter and efficiently learns system parameters online. We provide an interpretable regret bound for LARL with respect to the level of non-stationarity in the environment. LARL achieves sub-linear regret in this setting if the noise variance of the latent state process is sufficiently small with respect to $T$. Empirically, LARL outperforms various baseline methods in this non-stationary bandit problem.

## Contribution(s)

1. This paper introduces Latent AR LinUCB (LARL), an online algorithm designed for non-stationary MABs where the non-stationarity is due to a latent, auto-regressive (AR) state. LARL forms predictions of reward means by implicitly predicting the latent state using past rewards and actions. This strategy can be seen as an approximation of a steady-state Kalman filter with ground-truth system parameters.
   **Context:** The setting we consider is motivated by real-world applications where the non-stationary mechanism can be modeled by a latent state, but there is no budget for the non-stationarity. Existing approaches that consider similar settings rely on the latent state being discrete (Hong et al., 2020; Nelson et al., 2022) or require knowing the ground truth parameters or quality historical data to recover parameters (Liu et al., 2023; Chen et al., 2024).
2. We present an interpretable regret bound for LARL against the dynamic oracle. The regret bound allows practitioners to interpret the performance of LARL with respect to the level of non-stationarity in the environment and the complexity of learning parameters online.
   **Context:** Sub-linear regret with respect to the dynamic oracle is only possible in environments with a budget for non-stationarity that is sub-linear in $T$. For example, Besbes et al. (2014) assume a finite constant (variation budget) of how much the mean rewards can change over time and Garivier & Moulines (2011) assume a finite number of changes to the mean reward. We show that in our setting, LARL achieves sub-linear regret if the noise variance on the latent state process is sufficiently small with respect to $T$.
3. We demonstrate that LARL can outperform (achieve lower regret) against various stationary and non-stationary baselines in the non-stationary bandit environment where reward means change due to a latent AR state.
   **Context:** We consider cumulative regret across time and pairwise comparisons of methods in terms of total cumulative regret. To offer a fair comparison, baseline methods were only considered if they implemented an online learning strategy. For large values of $k$, the performance of LARL approaches the performance of baseline methods because the algorithm needs to fit more parameters, and thus requires more data to learn effectively.

# Hierarchical Multi-agent Reinforcement Learning for Cyber Network Defense

**Aditya Vikram Singh, Ethan Rathbun, Emma Graham, Lisa Oakley, Simona Boboila, Peter Chin, Alina Oprea**

**Keywords:** Multi-agent reinforcement learning, Cybersecurity, Deep reinforcement learning, Hierarchical reinforcement learning

## Summary

Recent advances in multi-agent reinforcement learning (MARL) have created opportunities to solve complex real-world tasks. Cybersecurity is a notable application area, where defending networks against sophisticated adversaries remains a challenging task typically performed by teams of security operators. In this work, we explore novel MARL strategies for building autonomous cyber network defenses that address challenges such as large policy spaces, partial observability, and stealthy, deceptive adversarial strategies. To facilitate efficient and generalized learning, we propose a hierarchical Proximal Policy Optimization (PPO) architecture that decomposes the cyber defense task into specific sub-tasks like network investigation and host recovery. Our approach involves training sub-policies for each sub-task using PPO enhanced with domain expertise. These sub-policies are then leveraged by a master defense policy that coordinates their selection to solve complex network defense tasks. Furthermore, the sub-policies can be fine-tuned and transferred with minimal cost to defend against shifts in adversarial behavior or changes in network settings. We conduct extensive experiments using CybORG Cage 4, the state-of-the-art MARL environment for cyber defense. Comparisons with multiple baselines across different adversaries show that our hierarchical learning approach achieves top performance in terms of convergence speed, episodic return, and several interpretable metrics relevant to cybersecurity, including the fraction of clean machines on the network, precision, and false positives.

## Contribution(s)

1. A scalable hierarchical multi-agent reinforcement learning method for cyber defense that decomposes the complex cyber defense task into multiple sub-tasks.
   **Context:** Prior work uses hierarchical MARL in other domains such as multi-robot learning, while current RL-based methods in the cyber defense domain are single agent.
2. A design guided by cybersecurity domain expertise to enhance the RL agents' observation space and facilitate learning of better policies.
   **Context:** Prior work on RL cyber defense uses the observation space provided by a cyber environment such as CybORG, without expanding it.
3. Defensive strategies that transfer either directly or via fine-tuning against a range of deceptive, stealthy adversaries in the CybORG CAGE 4 cyber environment.
   **Context:** We show that the proposed H-MARL methods generalize to three types of stealthy adversarial agents, besides the default red agent in CybORG CAGE 4, and we also demonstrate transferability to new red agents after fine-tuning.
4. Definition and analysis of multiple interpretable metrics for providing insights to security operators on the developed defenses.
   **Context:** Prior work in RL for cyber defense mainly analyzes the cumulative return, but does not discuss interpretable metrics, which are very relevant to security operators.

# A Finite-Time Analysis of Distributed Q-Learning

**Han-Dong Lim, Donghwan Lee**

**Keywords:** Q-learning, multi-agent reinforcement learning, distributed Q-learning

## Summary

Multi-agent reinforcement learning (MARL) has witnessed a remarkable surge in interest, fueled by the empirical success achieved in applications of single-agent reinforcement learning (RL). In this study, we consider a distributed Q-learning scenario, wherein a number of agents cooperatively solve a sequential decision making problem without access to the central reward function which is an average of the local rewards. In particular, we study finite-time analysis of a distributed Q-learning algorithm, and provide a new sample complexity result of $\tilde{\mathcal{O}}\left(\max\left\{\frac{1}{\epsilon^2}\frac{t_{\text{mix}}}{(1-\gamma)^6 d_{\min}^4}, \frac{1}{\epsilon}\frac{\sqrt{|\mathcal{S}||\mathcal{A}|}}{(1-\sigma_2(\boldsymbol{W}))(1-\gamma)^4 d_{\min}^3}\right\}\right)$ under tabular lookup setting for Markovian observation model.

## Contribution(s)

1. We provide a new sample complexity result for distributed Q-learning proposed in Kar et al. (2013). The analysis relies on construction on novel inequalities on the iterate of the averaged system.
   **Context:** The analysis of distributed Q-learning presented in Kar et al. (2013) is limited to asymptotic results. In contrast to Heredia et al. (2020); Zeng et al. (2022b), our approach relies on milder assumptions. The aforementioned works require additional strong assumptions, which may not hold even in the tabular setup. Furthermore, Wang et al. (2022) provided a version of distributed Q-learning which requires the communication process to occur after the update scheme. In contrast, our model allows the TD-error to be computed concurrently with the communication process.

# The Confusing Instance Principle for Online Linear Quadratic Control

**Waris Radji, Odalric-Ambrym Maillard**

**Keywords:** Model-based, linear quadratic regulator, exploration, minimum empirical divergence.

## Summary

We revisit the problem of controlling linear systems with quadratic cost under unknown dynamics within model-based reinforcement learning. Traditional methods like Optimism in the Face of Uncertainty and Thompson Sampling, rooted in multi-armed bandits (MABs), face practical limitations. In contrast, we propose an alternative based on the *Confusing Instance* (CI) principle, which underpins regret lower bounds in MABs and discrete Markov Decision Processes (MDPs) and is central to the *Minimum Empirical Divergence* (MED) family of algorithms, known for their asymptotic optimality in various settings. By leveraging the structure of LQR policies along with sensitivity and stability analysis, we develop `MED-LQ`. This novel control strategy extends CI and MED principles beyond small-scale settings.

Our work addresses a crucial research gap by exploring whether the CI principle can improve exploration strategies in continuous MDPs. While the exploration-exploitation dilemma is well understood in discrete settings, the curse of dimensionality makes this challenge significantly harder in continuous spaces. `MED-LQ` overcomes these challenges by efficiently searching for confusing instances through rank-one and entry-wise perturbations while avoiding intractable confidence bounds. Benchmarks on a comprehensive control suite demonstrate that `MED-LQ` achieves competitive performance across various scenarios, establishing foundations for a fresh perspective on exploration in continuous MDPs and opening new avenues for structured exploration in complex control problems.

## Contribution(s)

1. We formulate the Confusing Instance (CI) principle as an optimization problem in the LQR setting, extending this concept beyond MABs and discrete MDPs for the first time.
   **Context:** The CI principle has previously been applied only in discrete settings, primarily in multi-armed bandits and tabular MDPs (Honda & Takemura, 2010; 2015; Pesquerel & Maillard, 2022; Balagopalan & Jun, 2024).
2. We develop `MED-LQ`, a novel control strategy that implements the Minimum Empirical Divergence (MED) framework for online LQR, and show his numerical competitiveness.
   **Context:** Prior work established MED algorithms in discrete MDPs settings, with `IMED-RL` for ergodic case (Pesquerel & Maillard, 2022) and `IMED-KD` for the communicating case (Saber et al., 2024).
3. We develop a novel computational approach for building confusing instances in continuous systems through sensitivity analysis of rank-one perturbations.
   **Context:** Prior work limited confusing instances to discrete settings and linear bandits. Our sensitivity analysis for continuous control systems represents the first extension of this principle to linear dynamical systems.
4. We introduce `linquax`, a library for efficient research in online LQR problems, built with JAX to leverage automatic differentiation and provide GPU/TPU compatibility.
   **Context:** Prior to our work, no *modern* open-source library existed specifically for online LQR, creating a significant barrier to reproducible research in this domain.

# Drive Fast, Learn Faster: On-Board RL for High Performance Autonomous Racing

**Benedict Hildisch, Edoardo Ghignone, Nicolas Baumann, Cheng Hu, Andrea Carron, Michele Magno**

**Keywords:** Autonomous Racing, Physical Robot Learning,

## Summary

Autonomous racing presents unique challenges due to its non-linear dynamics, the high speed involved, and the critical need for real-time decision-making under dynamic and unpredictable conditions. Most traditional Reinforcement Learning (RL) approaches rely on extensive simulation-based pre-training, which faces crucial challenges in transfer effectively to real-world environments. This paper introduces a robust on-board RL framework for autonomous racing, designed to eliminate the dependency on simulation-based pre-training enabling direct real-world adaptation. The proposed system introduces a refined Soft Actor-Critic (SAC) algorithm, leveraging a residual RL structure to enhance classical controllers in real-time by integrating multi-step Temporal-Difference (TD) learning, an asynchronous training pipeline, and Heuristic Delayed Reward Adjustment (HDRA) to improve sample efficiency and training stability. The framework is validated through extensive experiments on the F1TENTH racing platform, where the residual RL controller consistently outperforms the baseline controllers and achieves up to an 11.5 % reduction in lap times compared to the State-of-the-Art (SotA) with only 20 min of training. Additionally, an End-to-End (E2E) RL controller trained without a baseline controller surpasses the previous best results with sustained on-track learning.

## Contribution(s)

1. This paper presents an efficient RL architecture, based on the residual policy structure, that enables uninterrupted on-board learning at an unprecedented level of performance. The architecture is compared to an E2E structure, showing how the residual policy greatly increases efficiency by simplifying the control task.
   **Context:** The work by Stachowicz et al. (2023) presented a high-speed driving solution with on-vehicle data only but failed to compare with SotA traditional methods for racing, exhibiting a performance gap. We show how a residual structure, previously used by Trumpp et al. (2023) in simulation, can be used to efficiently train in reality only and overtake classical SotA performance.
2. This paper compares different baseline controllers with the residual RL policy, showing how the proposed architecture is easily adaptable and improves on all the methods. The fastest agent in the comparison eventually turns out to be the fastest F1TENTH algorithm when compared to previous best results.
   **Context:** The work by Trumpp et al. (2023) only used Pure Pursuit (PP) as a baseline controller, while we show that the residual architecture can also adapt to different ones.
3. This paper provides an ablation study in the F1TENTH simulation environment of the different RL architectural choices, showing how the different parts affect efficiency and training consistency.
   **Context:** Some of the architectural choices were motivated by previous ablation studies (multi-step TD by Barth-Maron et al. (2018), asynchronous training by Yuan & Mahmood (2022)). Here the full combination is tested in the Autonomous Racing (AR) environment. Furthermore, the introduction of the novel HDRA is analyzed.
4. This paper provides a generalization study showing how the residual policy behaves with a different type of track and tires in a zero-shot and few-shot setup.
   **Context:** None

# Towards Large Language Models that Benefit for All: Benchmarking Group Fairness in Reward Models

**Kefan Song, Jin Yao, Runnan Jiang, Rohan Chandra, Shangtong Zhang**

**Keywords:** Group Fairness, Large Language Models, Reward Modeling, Reinforcement Learning from Human Feedback, Algorithmic Fairness.

## Summary

As Large Language Models (LLMs) become increasingly powerful and accessible to human users, ensuring fairness across diverse demographic groups, i.e., group fairness, is a critical ethical concern. However, current fairness and bias research in LLMs is limited in two aspects. First, compared to traditional group fairness in machine learning classification, it requires that the non-sensitive attributes, in this case, the question in the user prompts, be the same across different groups. In many practical scenarios, different groups, however, may prefer different questions and this requirement becomes impractical. Second, it evaluates group fairness only for the LLM's final output without identifying the source of possible bias. Namely, the bias in LLM's output can result from both the pretraining and the finetuning. For finetuning, the bias can result from both the RLHF procedure and the learned reward model. Arguably, evaluating the group fairness of each component in the LLM pipeline could help develop better methods to mitigate the possible bias. Recognizing those two limitations, this work benchmarks the group fairness of learned reward models. By using expert-written text from arXiv, we are able to benchmark the group fairness of reward models without requiring the same question in the user prompts across different demographic groups. Surprisingly, our results demonstrate that all the evaluated reward models (e.g., Nemotron-4-340B-Reward, ArmoRM-Llama3-8B-v0.1, and GRM-llama3-8B-sftreg) exhibit statistically significant group unfairness. We also observed that top-performing reward models (w.r.t. canonical performance metrics) tend to demonstrate better group fairness.

## Contribution(s)

1. We introduce a new problem of group fairness in reward models for LLMs, bridging a gap between algorithmic fairness methods and fairness research in LLMs.
   **Context:** Prior works (Lu et al., 2020; Garimella et al., 2022; Venkit et al., 2023; Bi et al., 2023) on LLM fairness predominantly addresses biased or harmful language in model outputs rather than unfairness within reward models.
2. We propose an evaluation methodology for group fairness that leverages a newly curated dataset derived from arXiv metadata.
   **Context:** None
3. We benchmark eight top-performing reward models from RewardBench (Lambert et al., 2024) and show that all exhibit statistically significant group unfairness.
   **Context:** None
4. We demonstrate that reward models with higher canonical performance metrics also tend to exhibit better group fairness, suggesting a possible link between overall model quality and fairness.
   **Context:** None

# Pure Exploration for Constrained Best Mixed Arm Identification with a Fixed Budget

**Dengwang Tang, Rahul Jain, Ashutosh Nayyar, Pierluigi Nuzzo**

**Keywords:** Constrained mixed arm identification, constrained bandit problem.

## Summary

We introduce the constrained best mixed arm identification (CBMAI) problem under unknown reward and costs wherein there are $K$ arms, each of which is associated with a reward and multiple cost attributes. These are random, and come from distributions with unknown means. The best mixed arm is a probability distribution over a subset of the $K$ arms that maximizes the expected reward while satisfying the expected cost constraints. We are specifically interested in a pure exploration problem under a fixed sampling budget with the goal of identifying the *support of the best mixed arm*. We propose a novel, parameter-free algorithm, called the Score Function-based Successive Reject (SFSR) algorithm, that combines the classical successive reject framework with a novel rejection criteria using a score function based on linear programming theory. We establish a performance guarantee for our algorithm by providing a theoretical upper bound on the probability of mis-identification of the support of the best mixed arm and show that it decays exponentially in the budget $N$ and some constants that characterize the hardness of the problem instance. We also develop an information-theoretic lower bound on the error probability that shows that these constants appropriately characterize the problem difficulty. We validate this empirically on a number of problem instances.

## Contribution(s)

1. This paper provides a novel, parameter-free algorithm that identifies the optimal support of the best mixed arm for a constrained best arm identification problem with a fixed sampling budget. We establish a performance guarantee for our algorithm in the form of an exponentially decaying, instance-dependent error upper bound.
   **Context:** Prior work has considered the best arm identification problem with *known costs* and/or with only *deterministic* arms allowed. However, we consider unknown costs and allow randomized (i.e., mixed) arms since deterministic arms may be suboptimal, or simply not meet all the constraints.

# Quantitative Resilience Modeling for Autonomous Cyber Defense

**Xavier Cadet, Simona Boboila, Edward Koh, Peter Chin, Alina Oprea**

**Keywords:** cyber resilience, reinforcement learning, evaluation metrics, operational cost, autonomous cyber defense.

## Summary

Cyber resilience is the ability of a system to recover from an attack with minimal impact on system operations. However, characterizing a network's resilience under a cyber attack is challenging, as there are no formal definitions of resilience applicable to diverse network topologies and attack patterns. In this work, we propose a quantifiable formulation of resilience that considers multiple defender operational goals, the criticality of various network resources for daily operations, and provides interpretability to security operators about their system's resilience under attack. We evaluate our approach within the CybORG environment, a reinforcement learning (RL) framework for autonomous cyber defense, analyzing trade-offs between resilience, costs, and prioritization of operational goals. Furthermore, we introduce methods to aggregate resilience metrics across time-variable attack patterns and multiple network topologies, comprehensively characterizing system resilience. Using insights gained from our resilience metrics, we design RL autonomous defensive agents and compare them against several heuristic baselines, showing that proactive network hardening techniques and prompt recovery of compromised machines are critical for effective cyber defenses.

## Contribution(s)

1. Formulation of a quantifiable resilience metric for autonomous cyber defense. The proposed metric captures the temporal evolution of system resilience as the attack progresses.
   **Context:** Prior work on resilience in the cyber defense domain are mostly qualitative discussions about generic system functionality in time (Huang et al., 2022; Zhao et al., 2025; Ligo et al., 2021; Linkov et al., 2023; Kott & Linkov, 2018; Fleming et al., 2021).
2. The proposed metric allows defenders to prioritize different objectives, such as confidentiality, availability, and integrity, and certain services, according to their operational goals.
   **Context:** Prior work discussing operational goals of confidentiality, integrity and availability in cyber environments is not formulating, evaluating or prioritizing them in the context of network resilience (Wiebe et al., 2023).
3. We develop new PPO-based defender agents that are trained to be proactive and to react quickly to attacks in the network. We demonstrate how these characteristics increase the resilience of a system under attack using the CybORG environment (Standen et al., 2021).
   **Context:** Prior work on autonomous cyber defense has not studied the impact of proactive and reactive defense characteristics on system resilience (Wiebe et al., 2023; Standen et al., 2021; Hammar et al., 2024).
4. We show how our resilience metric can be aggregated over multiple attack patterns and multiple network topologies to provide a comprehensive evaluation of the resilience of the system across various settings.
   **Context:** Weisman et al. (2025) presents results for autonomous vehicles, where markers of resilience like fuel efficiency are averaged over multiple runs. Our analysis studies additional levels of aggregation such as clustering of resilience evolution patterns.