Novelty Assessment Report

Paper: EXPO: Stable Reinforcement Learning with Expressive Policies
PDF URL: https://openreview.net/pdf?id=aFjSjkB6CV
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2026-01-01

Abstract

We study the problem of training and fine-tuning expressive policies with online reinforcement learning (RL) given an offline dataset. Training expressive policy classes with online RL present a unique challenge of stable value maximization. Unlike simpler Gaussian policies commonly used in online RL, expressive policies like diffusion and flow-matching policies are parameterized by a long denoising chain, which hinders stable gradient propagation from actions to policy parameters when optimizing against some value function. Our key insight is that we can address stable value maximization by avoiding direct optimization over value with the expressive policy and instead construct an on-the-fly RL policy to maximize Q-value. We propose Expressive Policy Optimization (EXPO), a sample-efficient online RL algorithm that utilizes an on-the-fly policy to maximize value with two parameterized policies -- a larger expressive base policy trained with a stable imitation learning objective and a light-weight Gaussian edit policy that edits the actions sampled from the base policy toward a higher value distribution. The on-the-fly policy optimizes the actions from the base policy with the learned edit policy and chooses the value maximizing action from the base and edited actions for both sampling and temporal-difference (TD) backup. Our approach yields up to 2-3x improvement in sample efficiency on average over prior methods both in the setting of fine-tuning a pretrained policy given offline data and in leveraging offline data to train online.

Core Task Landscape

This paper addresses: **Online Reinforcement Learning with Expressive Policy Classes**

A total of **50 papers** were analyzed and organized into a taxonomy with **33 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Expressive Policy Architectures and Representations**
- **Policy Optimization Algorithms and Training Frameworks**
- **State Representation and Observation Learning**
- **Multi-Agent and Interactive Decision Making**
- **Robustness and Adaptation in Sequential Decision Making**
- **Human-Centered and Preference-Based Learning**
- **Imitation and Demonstration-Based Learning**
- **Exploration and Sample Efficiency**
- **Offline Reinforcement Learning and Policy Evaluation**
- **Hierarchical and Structured Policy Representations**
- ... and 1 more categories

Complete Taxonomy Tree

- Online Reinforcement Learning with Expressive Policy Classes Survey Taxonomy
- Expressive Policy Architectures and Representations
  - Diffusion-Based Policy Learning
  - Online Diffusion Policy Training and Optimization ★ (5 papers)
    - [0] EXPO: Stable Reinforcement Learning with Expressive Policies (Anon et al., 2026) View paper
    - [3] Diffusion-based reinforcement learning via q-weighted variational policy optimization (Shutong Ding, 2024) View paper
    - [28] Efficient Online Reinforcement Learning for Diffusion Policy (Ma, 2025) View paper
    - [30] GoRL: An Algorithm-Agnostic Framework for Online Reinforcement Learning with Generative Policies (Chubin Zhang, 2025) View paper
    - [45] D2 Actor Critic: Diffusion Actor Meets Distributional Critic (Zhang Lun-jun, 2025) View paper
  - Offline and Theoretical Diffusion Policy Work (3 papers)
    - [26] Dynamic Rank Adjustment in Diffusion Policies for Efficient and Flexible Training (Xiatao Sun, 2025) View paper
    - [27] Policy agnostic rl: Offline rl and online rl fine-tuning of any class and backbone (Mark, 2024) View paper
    - [32] Policy representation via diffusion probability model for reinforcement learning (Yang, 2023) View paper
  - Flow-Based and Generative Policy Methods (2 papers)
  - [13] SAC Flow: Sample-Efficient Reinforcement Learning of Flow-Based Policies via Velocity-Reparameterized Sequential Modeling (Zhang Yi-xian, 2025) View paper
  - [29] FM-IRL: Flow-Matching for Reward Modeling and Policy Regularization in Reinforcement Learning (Wan ZhengLin, 2025) View paper
  - Maximum Entropy and Multimodal Policy Learning (2 papers)
  - [2] Maximum entropy reinforcement learning with diffusion policy (Dong, 2025) View paper

    - [31] DIME: Unifying Diffusion Models and Maximum Entropy Reinforcement Learning for Expressive Policy Representation (Ningkai, 2025) View paper
    - Discrete and Combinatorial Action Space Policies (1 papers)
    - [5] Reinforcement Learning with Discrete Diffusion Policies for Combinatorial Action Spaces (Ma, 2025) View paper
- Policy Optimization Algorithms and Training Frameworks
    - Policy Gradient and Actor-Critic Methods (2 papers)
    - [43] Natural actor-critic for robust reinforcement learning with function approximation (Zhou, 2023) View paper
    - [50] Trust region policy optimization (John Schulman, 2015) View paper
    - Value-Based and Q-Function Methods (2 papers)
    - [12] Online reinforcement learning via sparse Gaussian mixture model Q-functions (Minh Vu, 2025) View paper
    - [21] Inverse Policy Evaluation for Value-based Sequential Decision-making (Chan, 2020) View paper
    - Mixed Policy Scope and Contextual Decision Making (1 papers)
    - [1] Online reinforcement learning for mixed policy scopes (J Zhang, 2022) View paper
- State Representation and Observation Learning
    - Latent State and Predictive State Representations (2 papers)
    - [9] PAC Reinforcement Learning for Predictive State Representations (Zhan WenHao, 2022) View paper
    - [48] Provably efficient rl with rich observations via latent state decoding (Du, 2019) View paper
    - Action Representation and Contextualized Embeddings (2 papers)
    - [4] Learning Contextualized Action Representations in Sequential Decision Making for Adversarial Malware Optimization (Reza Ebrahimi, 2025) View paper
    - [6] Personalization in human-robot interaction through preference-based action representation learning (Ruiqi Wang, 2025) View paper
    - Memory-Augmented and Episodic Representations (1 papers)
    - [7] A scalable reinforcement learning framework inspired by hippocampal memory mechanisms for efficient contextual and sequential decision making (Hamed Poursiami, 2025) View paper
- Multi-Agent and Interactive Decision Making
    - Multi-Agent Representation and Transfer Learning (2 papers)
    - [10] Multi-Task Multi-Agent Reinforcement Learning With Interaction and Task Representations (Chao Li, 2024) View paper
    - [19] Enabling Multi-Agent Transfer Reinforcement Learning via Scenario Independent Representation (Ayesha Siddika Nipu, 2023) View paper
    - Interaction-Aware Driving and Coordination (2 papers)
    - [17] Modeling Interaction-Aware Driving Behavior using Graph-Based Representations and Multi-Agent Reinforcement Learning (Fabian Konstantinidis, 2023) View paper
    - [23] A Preference-Based Multi-Agent Federated Reinforcement Learning Algorithm Framework for Trustworthy Interactive Urban Autonomous Driving (Sikai Lu, 2025) View paper
    - Multi-Agent Coordination in Specialized Domains (2 papers)
    - [22] A Neuro-Symbolic Approach to Multi-Agent RL for Interpretability and Probabilistic Decision Making (Subramanian Chitra, 2024) View paper
    - [25] Enhancing Grid-Interactive Buildings Demand Response: Sequential Update-Based Multiagent Deep Reinforcement Learning Approach (Yinghua Han, 2024) View paper
    - Latent Strategy and Influence Learning (1 papers)
    - [37] Learning latent representations to influence multi-agent interaction (Xie, 2021) View paper
- Robustness and Adaptation in Sequential Decision Making
    - Robust Policy Learning Under Uncertainty (2 papers)
    - [14] Robust Learning and Evaluation in Sequential Decision Making (Keramati, 2021) View paper
    - [16] Action Robust Reinforcement Learning with Highly Expressive Policy (SeongIn Kim, 2024) View paper
    - Anomaly Detection in Decision Sequences (1 papers)
    - [8] OIL-AD: An Anomaly Detection Framework for Sequential Decision Sequences (Wang Chen, 2024) View paper
    - Nonstochastic and Adversarial Environments (1 papers)
    - [47] Online nonstochastic model-free reinforcement learning (Ghai, 2023) View paper
- Human-Centered and Preference-Based Learning
    - Preference-Based Reinforcement Learning (1 papers)
    - [35] LAPP: Large Language Model Feedback for Preference-Driven Reinforcement Learning (Wei Xiao, 2025) View paper
    - Online Behavior Modification and User Control (1 papers)
    - [20] Online Behavior Modification for Expressive User Control of RL-Trained Robots (Isaac Sheidlower, 2024) View paper
- Imitation and Demonstration-Based Learning
    - Behavior Tree and Hierarchical Task Learning (1 papers)
    - [36] Learning behavior trees from demonstration (Kevin French, 2019) View paper
    - Interaction and Manipulation Learning from Demonstration (1 papers)
    - [18] Learning Manipulation by Predicting Interaction (Jia Zeng, 2024) View paper
- Exploration and Sample Efficiency
    - Randomized and Function Approximation-Based Exploration (1 papers)
    - [24] Randomized exploration for reinforcement learning with multinomial logistic function approximation (Wooseong Cho, 2024) View paper
    - Workflow-Guided and Constrained Exploration (1 papers)
    - [15] Reinforcement learning on web interfaces using workflow-guided exploration (Liu, 2018) View paper
    - Reset-Based and Complexity-Measure-Driven Online RL (2 papers)
    - [41] The power of resets in online reinforcement learning (Dylan Foster, 2024) View paper
    - [49] GEC: A Unified Framework for Interactive Decision Making in MDP, POMDP, and Beyond (Zhong Han, 2022) View paper
- Offline Reinforcement Learning and Policy Evaluation
    - Offline Policy Evaluation and Estimation (1 papers)
    - [44] Opera: Automatic offline policy evaluation with re-weighted aggregates of multiple estimators (Anirudhan Badrinath, 2024) View paper

- Offline Policy Learning with Function Approximation (1 papers)
- [38] Revisiting the linear-programming framework for offline rl with general function approximation (Asuman Ozdaglar, 2023) View paper

- Hierarchical and Structured Policy Representations
  - Tree-Structured Policy Gradient and Recommendation (2 papers)
  - [42] Efficient Tree Policy with Attention-Based State Representation for Interactive Recommendation (Longxiang Shi, 2023) View paper
  - [46] Large-scale Interactive Recommendation with Tree-structured Policy Gradient (Haokun Chen, 2018) View paper
  - Decision Tree Function Approximation (1 papers)
  - [39] Fuzzy decision tree function approximation in reinforcement learning (Hitesh Shah, 2010) View paper
- Application-Specific Decision Making
  - Robotic Assembly and Sequential Planning (1 papers)
  - [11] Optimal Robotic Assembly Sequence Planning (ORASP): A Sequential Decision-Making Approach (Kartik Nagpal, 2024) View paper
  - Embodied Agents and Abstract-Concrete Reasoning (1 papers)
  - [33] Alfworld: Aligning text and embodied environments for interactive learning (Mohit Shridhar, 2020) View paper
  - Language-Guided Planning and Adaptive Representations (1 papers)
  - [34] Learning adaptive planning representations with natural language guidance (Wong, 2023) View paper
  - Multi-Sensory Object Property Learning (1 papers)
  - [40] MOSAIC: Learning Unified Multi-Sensory Object Property Representations for Robot Learning via Interactive Perception (Gyan Tatiya, 2023) View paper

Narrative

Core task: online reinforcement learning with expressive policy classes. The field encompasses a broad spectrum of approaches for learning sequential decision-making policies that can represent complex behaviors while training directly from interaction. The taxonomy organizes this landscape into several major branches: Expressive Policy Architectures and Representations explores rich function approximators such as diffusion models and flow-based policies; Policy Optimization Algorithms and Training Frameworks addresses the core algorithmic machinery for updating these policies; State Representation and Observation Learning tackles how agents encode and process environmental information; Multi-Agent and Interactive Decision Making considers settings with multiple learners or dynamic environments; and branches like Exploration and Sample Efficiency, Robustness and Adaptation, and Human-Centered Learning address complementary challenges in data collection, generalization, and alignment with human preferences. Application-Specific Decision Making and Hierarchical Policy Representations further refine methods for particular domains or structured action spaces, while Offline Reinforcement Learning provides a contrasting paradigm that learns from fixed datasets rather than online interaction.

Within the Expressive Policy Architectures branch, diffusion-based policy learning has emerged as a particularly active area, leveraging generative modeling techniques to represent multimodal action distributions. Works such as Q-Weighted Diffusion[3] and Efficient Diffusion[28] explore how to integrate value-based guidance and computational efficiency into diffusion policy training, while MaxEnt Diffusion[2] and Discrete Diffusion[5] extend these ideas to maximum-entropy frameworks and discrete action spaces. EXPO[0] sits squarely in this cluster, focusing on online diffusion policy training and optimization—a setting that contrasts with many offline diffusion methods by emphasizing direct interaction and iterative policy updates. Compared to Q-Weighted Diffusion[3], which emphasizes value-function weighting for action selection, EXPO[0] appears to prioritize the online training dynamics and optimization strategies needed to make diffusion policies practical in interactive environments. This positioning highlights ongoing questions about how to balance the expressive power of diffusion models with the sample efficiency and stability requirements of online RL, a trade-off that remains central across the broader taxonomy.

Related Works in Same Category

The following **4 sibling papers** share the same taxonomy leaf node with the original paper:

1. Diffusion-based reinforcement learning via q-weighted variational policy optimization

Authors: Shutong Ding, Ke Hu, Kan Ren, Ye Shi, Jingya Wang, et al. (8 authors total) | **Year/Venue**: 2024 | **URL**: View paper

Abstract

â□¦ and policy loss in online RL, we â□¦ policy expressiveness of the diffusion model decreases with fewer diffusion steps. In our experiments, we find that not only does policy expressiveness â□¦

Relationship Analysis

Both papers belong to the same taxonomy category of online diffusion policy training and optimization, addressing the challenge of stable value maximization with expressive diffusion policies in online RL. They share the common goal of enabling effective online fine-tuning of diffusion policies while avoiding gradient pathologies through the long denoising chain. The key difference is that EXPO uses a two-policy approach with a base expressive policy trained via imitation learning and a lightweight Gaussian edit policy for value maximization, while QVPO proposes a Q-weighted variational lower bound objective that directly trains the diffusion policy with weighted samples and includes entropy regularization for exploration.

---

2. Efficient Online Reinforcement Learning for Diffusion Policy

Authors: Ma, Haitong, Chen, Tianyi, Haitong Ma, et al. (12 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Diffusion policies have achieved superior performance in imitation learning and offline reinforcement learning (RL) due to their rich expressiveness. However, the conventional diffusion training procedure requires samples from target distribution, which is impossible in online RL since we cannot sample from the optimal policy. Backpropagating policy gradient through the diffusion process incurs huge computational costs and instability, thus being expensive and not scalable. To enable efficient t...

Relationship Analysis

Both papers belong to the same taxonomy category focusing on stable online RL training of diffusion policies, addressing gradient pathologies in value-based optimization. They share the challenge of training expressive diffusion policies without unstable backpropagation through long denoising chains. However, EXPO avoids direct value optimization of the expressive policy by using a separate lightweight Gaussian edit policy and on-the-fly action selection, while the candidate paper (Efficient Online RL for Diffusion Policy) proposes Reweighted Score Matching (RSM) to generalize denoising score matching by reweighting the loss function, enabling direct diffusion policy training for policy mirror descent and max-entropy objectives without sampling from the optimal policy.

---

3. GoRL: An Algorithm-Agnostic Framework for Online Reinforcement Learning with Generative Policies

Authors: Chubin Zhang, Zhenglin Wan, Feng Chen, Xingrui Yu, Ivor Tsang, et al. (6 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Reinforcement learning (RL) faces a persistent tension: policies that are stable to optimize are often too simple to represent the multimodal action distributions needed for complex control. Gaussian policies provide tractable likelihoods and smooth gradients, but their unimodal form limits expressiveness. Conversely, generative policies based on diffusion or flow matching can model rich multimodal behaviors; however, in online RL, they are frequently unstable due to intractable likelihoods and ...

Relationship Analysis

Both papers belong to the same taxonomy category focused on stable online RL training of expressive policies, specifically addressing gradient pathologies in diffusion-based and generative policy learning. They overlap in their core motivation to enable stable value maximization with expressive policy classes during online RL, and both propose methods to avoid direct backpropagation through long generative chains. However, EXPO uses a two-policy architecture (base expressive policy + lightweight edit policy) with on-the-fly action selection, while GoRL employs a decoupled latent-space optimization approach with a conditional generative decoder and two-timescale updates.

---

4. D2 Actor Critic: Diffusion Actor Meets Distributional Critic

Authors: Zhang Lun-jun, Han, Shuo, Lunjun Zhang, Shuo Han, et al. (9 authors total) | **Year/Venue**: 2025 • Trans. Mach. Learn. Res. | **URL**: View paper

Abstract

We introduce D2AC, a new model-free reinforcement learning (RL) algorithm designed to train expressive diffusion policies online effectively. At its core is a policy improvement objective that avoids the high variance of typical policy gradients and the complexity of backpropagation through time. This stable learning process is critically enabled by our second contribution: a robust distributional critic, which we design through a fusion of distributional RL and clipped double Q-learning. The re...

Relationship Analysis

Both papers belong to the same taxonomy category focusing on online diffusion policy training and optimization, addressing the challenge of stable RL training with expressive policies. They overlap in their goal of enabling effective online RL with diffusion-based policies while maintaining training stability and sample efficiency. However, EXPO uses a two-policy approach (base expressive policy trained via imitation learning plus a lightweight Gaussian edit policy) with on-the-fly action selection, whereas D2AC introduces a distributional critic combined with a policy improvement objective that avoids backpropagation through the diffusion chain, representing fundamentally different architectural and optimization strategies.

Contributions Analysis

Overall novelty summary. The paper proposes EXPO, an algorithm for training expressive policies (specifically diffusion and flow-matching models) in online reinforcement learning settings. It resides in the 'Online Diffusion Policy Training and Optimization' leaf, which contains five papers total including the original work. This leaf sits within the broader 'Diffusion-Based Policy Learning' branch, which itself is part of 'Expressive Policy Architectures and Representations'. The taxonomy reveals this is a moderately populated research direction: the immediate leaf has four sibling papers, and the parent branch includes an additional three papers on offline and theoretical diffusion policy work, indicating sustained but not overwhelming activity in diffusion-based RL methods.

The taxonomy structure shows that EXPO's leaf is adjacent to 'Offline and Theoretical Diffusion Policy Work' (three papers) and sits alongside other expressive policy branches including 'Flow-Based and Generative Policy Methods' (two papers) and 'Maximum Entropy and Multimodal Policy Learning' (two papers). The scope note for EXPO's leaf emphasizes 'stable online RL training of diffusion policies addressing gradient pathologies and sample efficiency', while explicitly excluding offline-only methods. This boundary clarifies that EXPO targets the specific challenge of interactive learning with diffusion models, distinguishing it from offline behavioral cloning approaches and from flow-matching methods that use different generative architectures.

Among the three contributions analyzed, the literature search examined twenty-eight candidate papers total. For the core EXPO algorithm contribution, nine candidates were examined with zero refutable matches. The on-the-fly policy parameterization examined nine candidates (zero refutable), and the edit policy with distance constraint examined ten candidates (zero refutable). These statistics indicate that within the limited search scope—roughly thirty semantically similar papers—no prior work was identified that clearly overlaps with EXPO's specific combination of techniques. The absence of refutable candidates across all three contributions suggests the approach may occupy a relatively unexplored niche, though the search scale means this assessment is necessarily provisional.

Based on the limited literature search of approximately thirty candidates, EXPO appears to introduce a novel combination of mechanisms for online diffusion policy training. The taxonomy context reveals this work sits in an active but not saturated research area, with roughly a dozen papers across diffusion-based policy learning. The zero-refutation finding across all contributions should be interpreted cautiously given the search scope: it indicates no obvious prior work among top semantic matches, but does not constitute exhaustive verification of novelty across the entire field.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: EXPO algorithm for stable online RL with expressive policies

Description: EXPO is a novel online reinforcement learning algorithm designed to train and fine-tune expressive policy classes (such as diffusion or flow-matching policies) in a stable manner. The method avoids direct value optimization of the expressive policy by using a base policy trained with imitation learning and a lightweight Gaussian edit policy that refines actions toward higher Q-values, combined with an on-the-fly action selection mechanism.

This contribution was assessed against **9 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Dual-Actor Fine-Tuning of VLA Models: A Talk-and-Tweak Human-in-the-Loop Approach

URL: View paper

Brief Assessment

Dual-Actor VLA[62] focuses on human-in-the-loop fine-tuning of vision-language-action models for robotic manipulation using a dual-actor framework with language-based corrections, not on general online RL algorithms with expressive policies like diffusion or flow-matching.

---

2. Accelerating Deep Reinforcement Learning Using Human Demonstration Data Based on Dual Replay Buffer Management and Online Frame Skipping

URL: View paper

Brief Assessment

Dual Replay Buffer[66] focuses on managing human demonstration data through dual replay buffers and frame skipping techniques for Atari games, not on training expressive policies (diffusion/flow-matching) with dual parameterization and on-the-fly action selection mechanisms.

---

3. Knowledge-Level Consistency Reinforcement Learning: Dual-Fact Alignment for Long-Form Factuality

URL: View paper

Brief Assessment

Knowledge-Level Consistency[67] addresses factuality in LLM text generation through knowledge consistency rewards, not online RL algorithms for expressive policies in continuous control tasks.

---

4. A Two-Timescale Primal-Dual Framework for Reinforcement Learning via Online Dual Variable Guidance

URL: View paper

Brief Assessment

Two-Timescale Primal-Dual[65] focuses on primal-dual optimization for regularized MDPs with experience replay, not on training expressive policies (diffusion/flow-matching) with dual parameterization and edit policies.

---

5. A Distributed Primal-Dual Method for Constrained Multi-agent Reinforcement Learning with General Parameterization

URL: View paper

Brief Assessment

Distributed Primal-Dual[68] addresses multi-agent constrained RL with distributed coordination, not single-agent online RL with expressive policies like diffusion or flow-matching models.

---

6. Finite-Horizon Optimal Control for Nonlinear Multi-Input Systems With Online Adaptive Integral Reinforcement Learning

URL: View paper

Brief Assessment

Finite-Horizon Optimal[61] focuses on finite-horizon optimal control for multi-input nonlinear systems using integral reinforcement learning with dual neural networks, not on training expressive policies (diffusion/flow-matching) with a dual parameterization approach involving base and edit policies.

---

7. OMPO: A Unified Framework for RL under Policy and Dynamics Shifts

URL: View paper

Brief Assessment

OMPO[64] addresses policy and dynamics shifts through occupancy matching with a discriminator-based approach, while EXPO focuses on stable value maximization for expressive policies using dual parameterization (base + edit policy). These are fundamentally different technical approaches to different problems.

---

8. Reward Shaping via Diffusion Process in Reinforcement Learning

URL: View paper

Brief Assessment

Diffusion Reward Shaping[69] focuses on reward shaping via diffusion processes in RL from a stochastic thermodynamics perspective, not on dual-policy architectures for stable value maximization with expressive policies. The candidate does not address the specific challenge of training expressive policies (diffusion/flow-matching) with online RL through edit policies and on-the-fly action selection.

---

9. Reinforcement learning via gaussian processes with neural network dual kernels

URL: View paper

Brief Assessment

Gaussian Neural Kernels[63] focuses on applying Gaussian processes with neural network dual kernels to reinforcement learning problems, specifically the mountain-car problem. This is fundamentally different from EXPO's approach of using dual parameterization with a base expressive policy and edit policy for stable online RL fine-tuning.

Contribution 2: On-the-fly policy parameterization for value maximization

Description: The authors introduce an on-the-fly policy construction that performs value maximization without directly optimizing the expressive policy parameters. This policy samples actions from both the base expressive policy and an edit policy, then selects the highest Q-value action for both environment interaction and temporal-difference backup, enabling more stable and immediate reflection of Q-function changes.

This contribution was assessed against **9 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Combining policy gradient and Q-learning

URL: View paper

Brief Assessment

Policy Gradient Q-Learning[73] focuses on combining policy gradient with Q-learning through a fixed-point relationship between regularized policy gradient and Q-values, not on constructing an on-the-fly policy that samples from both base and edit policies for value maximization as in the original paper.

---

2. Decoupled policy actor-critic: Bridging pessimism and risk awareness in reinforcement learning

URL: View paper

Brief Assessment

Decoupled Policy[72] focuses on decoupling pessimistic and optimistic actors for TD learning versus exploration, not on constructing an on-the-fly policy that samples from both base and edit policies then selects highest Q-value actions for both interaction and backup as in the original paper.

---

3. Continuous-time reinforcement learning for optimal switching over multiple regimes

URL: View paper

Brief Assessment

Continuous-Time Switching[74] addresses optimal regime switching in continuous-time using entropy regularization and generator matrix control, not on-the-fly policy construction for discrete action selection in RL. The technical approaches are fundamentally different.

---

4. Meta-learning strategies through value maximization in neural networks

URL: View paper

Brief Assessment

Meta-Learning Strategies[75] focuses on meta-learning control signals for neural network training dynamics in supervised/reinforcement learning contexts, not on policy construction for action selection in RL environments. The candidate's value maximization operates over learning trajectories and hyperparameters, while the original paper addresses immediate action-level policy extraction.

---

5. Learning without gradients: multi-agent reinforcement learning approach to optimization

URL: View paper

Brief Assessment

Without Gradients[77] focuses on multi-agent RL for optimization without gradients in machine learning training, not on constructing on-the-fly policies for value maximization in standard RL tasks. The technical approaches are fundamentally different.

---

6. Pass@K Policy Optimization: Solving Harder Reinforcement Learning Problems

URL: View paper

Brief Assessment

Pass@K[70] focuses on optimizing pass@k performance through reward transformations in standard RL settings, not on constructing on-the-fly policies that combine base and edit policies for value maximization without direct policy gradient optimization.

---

7. Unraveling the Rainbow: can value-based methods schedule?

URL: View paper

Brief Assessment

Unraveling Rainbow[78] focuses on value-based methods (DQN, Rainbow) for job-shop scheduling problems, not on-the-fly policy construction for expressive policies. The candidate addresses discrete action spaces in combinatorial optimization, while the original addresses continuous control with expressive policies like diffusion models.

---

8. Learning in complex action spaces without policy gradients

URL: View paper

Brief Assessment

Without Policy Gradients[76] focuses on action-value methods using sampling-based arg max approximation and MLE for amortized maximization in Q-learning, not on constructing policies that combine base and edit policies for temporal-difference backup as in the original paper.

---

9. Inverse Reinforcement Learning with Explicit Policy Estimates

URL: View paper

Brief Assessment

Explicit Policy IRL[71] focuses on inverse reinforcement learning methods that estimate policies from expert demonstrations, not on direct value maximization in online RL. The candidate's on-the-fly policy construction is for reward inference from observed behavior, while the original contribution addresses online RL training of expressive policies without policy gradient optimization.

Contribution 3: Edit policy with distance constraint for action refinement

Description: The method introduces a small Gaussian edit policy that locally refines actions generated by the base expressive policy to maximize Q-values while maintaining proximity to the original actions through an edit distance constraint. This design allows the edit policy to solve a simpler local optimization problem, enabling efficient training with entropy regularization for exploration.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. A Train Cooperative Operation Optimization Method Considering Passenger Comfort based on Reinforcement Learning

URL: View paper

Brief Assessment

Train Cooperative[52] focuses on train speed curve optimization using TD3 for cooperative operation with jerk rate constraints. It does not address action refinement policies with distance constraints for Q-value maximization in the context of expressive policy optimization.

---

2. Adaptive auxiliary task selection for multitasking-assisted constrained multi-objective optimization

URL: View paper

Brief Assessment

Auxiliary Task[54] focuses on multitasking-assisted constrained multi-objective optimization with auxiliary task selection, not on action refinement policies for reinforcement learning with distance constraints.

---

3. Improved DQN-based Robot Path Planning Algorithm for Mobile Robots

URL: View paper

Brief Assessment

Improved DQN[57] focuses on robot path planning with distance-based rewards for navigation, not action refinement policies that edit base policy outputs with distance constraints for Q-value maximization in general RL settings.

---

4. Reinforcement learning with distance-based incentive/penalty (DIP) updates for highly constrained industrial control systems

URL: View paper

Brief Assessment

Distance-Based DIP[59] focuses on constrained industrial control with discrete/continuous actions using distance-based Q-value updates, not on refining actions from expressive policies for local Q-value optimization.

---

5. Flow-Based Policy for Online Reinforcement Learning

URL: View paper

Brief Assessment

Flow-Based Policy[51] uses flow-based generative models with Wasserstein-2 distance constraints for policy optimization, not a Gaussian edit policy that refines base policy actions with edit distance constraints as in the original paper.

---

6. Results in Engineering

URL: View paper

Brief Assessment

Results Engineering[60] provides insufficient context to assess novelty claims. The candidate's fragments mention distance constraints and q-value gradients but lack detail on the specific edit policy architecture, training objectives, or the local optimization framework described in the original paper.

---

7. Decision Boundary Estimation Using Reinforcement Learning for Complex Classification Problems

URL: View paper

Brief Assessment

Decision Boundary[55] focuses on finding decision boundary points for SVM classifiers in constrained systems, not on refining actions from a base policy for Q-value maximization in RL.

---

8. A Proportional Navigation Based Reinforcement Learning Approach: Efficiency Enhancements and Practical Applications

URL: View paper

Brief Assessment

Proportional Navigation[58] focuses on proportional navigation guidance for pursuer-target interception using Q-learning with reward shaping, not on edit policies with distance constraints for action refinement in expressive policy optimization.

---

9. A Q-Learning-Based Particle Swarm Optimization for Aircraft Routing and Scheduling in Airport Terminal Area

URL: View paper

Brief Assessment

Q-Learning PSO[56] applies Q-learning to particle swarm optimization for aircraft routing, not to action refinement in reinforcement learning policies. The distance constraints in Q-Learning PSO[56] refer to safe distances between aircraft, not edit distance constraints on policy actions.

---

10. Service-Oriented Segmented Trajectory Design for Low-altitude UAV-Assisted MEC Networks

URL: View paper

Brief Assessment

UAV MEC[53] focuses on trajectory design for UAV-assisted mobile edge computing networks with energy-based distance constraints, not reinforcement learning action refinement policies.

Appendix: Text Similarity Detection

No high-similarity text segments were detected across any compared papers.

References

- [0] EXPO: Stable Reinforcement Learning with Expressive Policies View paper
- [1] Online reinforcement learning for mixed policy scopes View paper
- [2] Maximum entropy reinforcement learning with diffusion policy View paper
- [3] Diffusion-based reinforcement learning via q-weighted variational policy optimization View paper
- [4] Learning Contextualized Action Representations in Sequential Decision Making for Adversarial Malware Optimization View paper
- [5] Reinforcement Learning with Discrete Diffusion Policies for Combinatorial Action Spaces View paper
- [6] Personalization in human-robot interaction through preference-based action representation learning View paper
- [7] A scalable reinforcement learning framework inspired by hippocampal memory mechanisms for efficient contextual and sequential decision making View paper
- [8] OIL-AD: An Anomaly Detection Framework for Sequential Decision Sequences View paper
- [9] PAC Reinforcement Learning for Predictive State Representations View paper
- [10] Multi-Task Multi-Agent Reinforcement Learning With Interaction and Task Representations View paper
- [11] Optimal Robotic Assembly Sequence Planning (ORASP): A Sequential Decision-Making Approach View paper
- [12] Online reinforcement learning via sparse Gaussian mixture model Q-functions View paper
- [13] SAC Flow: Sample-Efficient Reinforcement Learning of Flow-Based Policies via Velocity-Reparameterized Sequential Modeling View paper
- [14] Robust Learning and Evaluation in Sequential Decision Making View paper
- [15] Reinforcement learning on web interfaces using workflow-guided exploration View paper
- [16] Action Robust Reinforcement Learning with Highly Expressive Policy View paper
- [17] Modeling Interaction-Aware Driving Behavior using Graph-Based Representations and Multi-Agent Reinforcement Learning View paper
- [18] Learning Manipulation by Predicting Interaction View paper
- [19] Enabling Multi-Agent Transfer Reinforcement Learning via Scenario Independent Representation View paper
- [20] Online Behavior Modification for Expressive User Control of RL-Trained Robots View paper
- [21] Inverse Policy Evaluation for Value-based Sequential Decision-making View paper
- [22] A Neuro-Symbolic Approach to Multi-Agent RL for Interpretability and Probabilistic Decision Making View paper
- [23] A Preference-Based Multi-Agent Federated Reinforcement Learning Algorithm Framework for Trustworthy Interactive Urban Autonomous Driving View paper
- [24] Randomized exploration for reinforcement learning with multinomial logistic function approximation View paper
- [25] Enhancing Grid-Interactive Buildings Demand Response: Sequential Update-Based Multiagent Deep Reinforcement Learning Approach View paper

- [26] Dynamic Rank Adjustment in Diffusion Policies for Efficient and Flexible Training View paper
- [27] Policy agnostic rl: Offline rl and online rl fine-tuning of any class and backbone View paper
- [28] Efficient Online Reinforcement Learning for Diffusion Policy View paper
- [29] FM-IRL: Flow-Matching for Reward Modeling and Policy Regularization in Reinforcement Learning View paper
- [30] GoRL: An Algorithm-Agnostic Framework for Online Reinforcement Learning with Generative Policies View paper
- [31] DIME: Unifying Diffusion Models and Maximum Entropy Reinforcement Learning for Expressive Policy Representation View paper
- [32] Policy representation via diffusion probability model for reinforcement learning View paper
- [33] Alfworld: Aligning text and embodied environments for interactive learning View paper
- [34] Learning adaptive planning representations with natural language guidance View paper
- [35] LAPP: Large Language Model Feedback for Preference-Driven Reinforcement Learning View paper
- [36] Learning behavior trees from demonstration View paper
- [37] Learning latent representations to influence multi-agent interaction View paper
- [38] Revisiting the linear-programming framework for offline rl with general function approximation View paper
- [39] Fuzzy decision tree function approximation in reinforcement learning View paper
- [40] MOSAIC: Learning Unified Multi-Sensory Object Property Representations for Robot Learning via Interactive Perception View paper
- [41] The power of resets in online reinforcement learning View paper
- [42] Efficient Tree Policy with Attention-Based State Representation for Interactive Recommendation View paper
- [43] Natural actor-critic for robust reinforcement learning with function approximation View paper
- [44] Opera: Automatic offline policy evaluation with re-weighted aggregates of multiple estimators View paper
- [45] D2 Actor Critic: Diffusion Actor Meets Distributional Critic View paper
- [46] Large-scale Interactive Recommendation with Tree-structured Policy Gradient View paper
- [47] Online nonstochastic model-free reinforcement learning View paper
- [48] Provably efficient rl with rich observations via latent state decoding View paper
- [49] GEC: A Unified Framework for Interactive Decision Making in MDP, POMDP, and Beyond View paper
- [50] Trust region policy optimization View paper
- [51] Flow-Based Policy for Online Reinforcement Learning View paper
- [52] A Train Cooperative Operation Optimization Method Considering Passenger Comfort based on Reinforcement Learning View paper
- [53] Service-Oriented Segmented Trajectory Design for Low-altitude UAV-Assisted MEC Networks View paper
- [54] Adaptive auxiliary task selection for multitasking-assisted constrained multi-objective optimization View paper
- [55] Decision Boundary Estimation Using Reinforcement Learning for Complex Classification Problems View paper
- [56] A Q-Learning-Based Particle Swarm Optimization for Aircraft Routing and Scheduling in Airport Terminal Area View paper
- [57] Improved DQN-based Robot Path Planning Algorithm for Mobile Robots View paper
- [58] A Proportional Navigation Based Reinforcement Learning Approach: Efficiency Enhancements and Practical Applications View paper
- [59] Reinforcement learning with distance-based incentive/penalty (DIP) updates for highly constrained industrial control systems View paper
- [60] Results in Engineering View paper
- [61] Finite-Horizon Optimal Control for Nonlinear Multi-Input Systems With Online Adaptive Integral Reinforcement Learning View paper
- [62] Dual-Actor Fine-Tuning of VLA Models: A Talk-and-Tweak Human-in-the-Loop Approach View paper
- [63] Reinforcement learning via gaussian processes with neural network dual kernels View paper
- [64] OMPO: A Unified Framework for RL under Policy and Dynamics Shifts View paper
- [65] A Two-Timescale Primal-Dual Framework for Reinforcement Learning via Online Dual Variable Guidance View paper
- [66] Accelerating Deep Reinforcement Learning Using Human Demonstration Data Based on Dual Replay Buffer Management and Online Frame Skipping View paper
- [67] Knowledge-Level Consistency Reinforcement Learning: Dual-Fact Alignment for Long-Form Factuality View paper
- [68] A Distributed Primal-Dual Method for Constrained Multi-agent Reinforcement Learning with General Parameterization View paper
- [69] Reward Shaping via Diffusion Process in Reinforcement Learning View paper
- [70] Pass@K Policy Optimization: Solving Harder Reinforcement Learning Problems View paper
- [71] Inverse Reinforcement Learning with Explicit Policy Estimates View paper
- [72] Decoupled policy actor-critic: Bridging pessimism and risk awareness in reinforcement learning View paper
- [73] Combining policy gradient and Q-learning View paper
- [74] Continuous-time reinforcement learning for optimal switching over multiple regimes View paper
- [75] Meta-learning strategies through value maximization in neural networks View paper
- [76] Learning in complex action spaces without policy gradients View paper
- [77] Learning without gradients: multi-agent reinforcement learning approach to optimization View paper
- [78] Unraveling the Rainbow: can value-based methods schedule? View paper