Novelty Assessment Report

Paper: Relative Entropy Pathwise Policy Optimization

PDF URL: https://openreview.net/pdf?id=4vmm8mlHkS

Venue: ICLR 2026 Conference Submission

Year: 2026

Report Generated: 2026-01-07

Abstract

Score-function based methods for policy learning, such as REINFORCE and PPO, have delivered strong results in game-playing and robotics, yet their high variance often undermines training stability. Using pathwise policy gradients, i.e. computing a derivative by differentiating the objective function, alleviates the variance issues. However, they require an accurate action-conditioned value function, which is notoriously hard to learn without relying on replay buffers for reusing past off-policy data. We present an on-policy algorithm that trains Q-value models purely from on-policy trajectories, unlocking the possibility of using pathwise policy updates in the context of on-policy learning. We show how to combine stochastic policies for exploration with constrained updates for stable training, and evaluate important architectural components that stabilize value function learning. The result, Relative Entropy Pathwise Policy Optimization (REPPO), is an efficient on-policy algorithm that combines the stability of pathwise policy gradients with the simplicity and minimal memory footprint of standard on-policy learning. Compared to state-of-the-art on two standard GPU-parallelized benchmarks, REPPO provides strong empirical performance at superior sample efficiency, wall-clock time, memory footprint, and hyperparameter robustness.

Disclaimer

This report is **AI-GENERATED** using Large Language Models and WisPaper (a scholar search engine). It analyzes academic papers' tasks and contributions against retrieved prior work. While this system identifies **POTENTIAL** overlaps and novel directions, **ITS COVERAGE IS NOT EXHAUSTIVE AND JUDGMENTS ARE APPROXIMATE**. These results are intended to assist human reviewers and **SHOULD NOT** be relied upon as a definitive verdict on novelty.

Note that some papers exist in multiple, slightly different versions (e.g., with different titles or URLs). The system may retrieve several versions of the same underlying work. The current automated pipeline does not reliably align or distinguish these cases, so human reviewers will need to disambiguate them manually.

If you have any questions, please contact: mingzhang23@m.fudan.edu.cn

Core Task Landscape

This paper addresses: **On-Policy Reinforcement Learning with Pathwise Policy Gradients**

A total of **6 papers** were analyzed and organized into a taxonomy with **5 categories**.

Taxonomy Overview

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

- **Pathwise Gradient Estimation Methods**
- **Model-Based Policy Search with Gradient Stability**
- **Application Domains and Specialized Implementations**
- **Theoretical Foundations and Control-Theoretic Perspectives**

Complete Taxonomy Tree

- On-Policy Reinforcement Learning with Pathwise Policy Gradients Survey Taxonomy
- Pathwise Gradient Estimation Methods
  - On-Policy Pathwise Optimization ★ (2 papers)
  - [0] Relative Entropy Pathwise Policy Optimization (Anon et al., 2026) View paper
  - [1] Deep policy gradient methods without batch updates, target networks, or replay buffers (Alireza Azimi, 2024) View paper
  - Discrete Action Space Adaptations (1 papers)
  - [5] Deterministic Policy Optimization by Combining Pathwise and Score Function Estimators for Discrete Action Spaces (Levy, 2017) View paper
- Model-Based Policy Search with Gradient Stability (1 papers)
  - [3] PIPPS: Flexible Model-Based Policy Search Robust to the Curse of Chaos (Parmas, 2019) View paper
- Application Domains and Specialized Implementations (1 papers)
  - [2] Differentiable discrete event simulation for queuing network control (Che, 2024) View paper
- Theoretical Foundations and Control-Theoretic Perspectives (2 papers)
  - [4] A Differential and Pointwise Control Approach to Reinforcement Learning (MP Nguyen, 2025) View paper
  - [6] Control-based theories for credit assignment in neural networks and reinforcement learning (Meulemans, 2024) View paper

Narrative

Core task: on-policy reinforcement learning with pathwise policy gradients. This field centers on computing policy gradients by differentiating through entire trajectories or environment dynamics, rather than relying solely on score-function estimators. The taxonomy reveals four main branches. Pathwise Gradient Estimation Methods explore direct differentiation techniques, including on-policy pathwise optimization and strategies for handling discrete or stochastic transitions. Model-Based Policy Search with Gradient Stability investigates how learned or analytic models can provide stable gradient signals, often trading off sample efficiency against model fidelity. Application Domains and Specialized Implementations address deployment in robotics, control tasks, and other settings where pathwise gradients offer practical advantages. Theoretical Foundations and Control-Theoretic Perspectives ground these methods in optimization theory and classical control, clarifying convergence properties and connections to deterministic optimal control.

Recent work has intensified around making pathwise gradients practical in challenging scenarios. Some studies focus on discrete action spaces, where reparameterization is nontrivial—Deterministic Discrete Actions[5] and Differentiable Discrete Event[2] exemplify efforts to enable gradient flow despite discontinuities. Others emphasize on-policy stability and credit assignment: Deep Policy Without Batch[1] and Control Credit Assignment[6] tackle how to maintain low-variance updates without large replay buffers, while PIPPS[3] and Differential Pointwise Control[4] refine gradient estimation under model uncertainty. Relative Entropy Pathwise[0] sits within the on-policy pathwise optimization cluster, sharing with Deep Policy Without Batch[1] an emphasis on direct policy updates but distinguishing itself through a relative-entropy regularization framework that aims to balance exploration and gradient stability. Compared to PIPPS[3], which leverages model-based rollouts, Relative Entropy Pathwise[0] operates more directly on sampled trajectories, highlighting an ongoing tension between sample efficiency and the complexity of maintaining differentiable environment models.

Related Works in Same Category

The following **1 sibling papers** share the same taxonomy leaf node with the original paper:

1. Deep policy gradient methods without batch updates, target networks, or replay buffers

Authors: Alireza Azimi, Colin Bellinger, mohamed elsayed, JiaMin He, A Mahmood, et al. (8 authors total) | **Year/Venue**: 2024 | **URL**: View paper

Abstract

â¦ estimating the gradient, the reparameterization gradient (RG) estimator or the pathwise gradient â¦ , which is a commonly used technique in on-policy RL algorithms such as PPO to attain â¦

Relationship Analysis

Both papers belong to the On-Policy Pathwise Optimization category, focusing on combining pathwise gradients with on-policy learning without replay buffers. They overlap in using pathwise policy gradients (reparameterization gradients) for on-policy learning and addressing the challenge of learning accurate value functions from on-policy data alone. However, REPPO uses multi-step TD-λ targets with maximum entropy formulation and KL-constrained updates for continuous control, while AVG (Action Value Gradient) focuses on incremental learning with single-step updates, normalization techniques, and demonstrates real robot applications, positioning itself as a resource-constrained incremental method rather than a batch on-policy approach.

Contributions Analysis

Overall novelty summary. The paper proposes an on-policy reinforcement learning algorithm that applies pathwise policy gradients without replay buffers, combining stochastic exploration with constrained updates and architectural innovations for stable value learning. It resides in the 'On-Policy Pathwise Optimization' leaf, which contains only two papers total. This sparse taxonomy leaf suggests the approach targets a relatively underexplored niche: most pathwise gradient methods either rely on off-policy data or operate in model-based settings, whereas this work pursues pure on-policy trajectories for gradient computation.

The taxonomy reveals that neighboring research directions include discrete action space adaptations, model-based policy search addressing gradient instability, and theoretical control-theoretic frameworks. The paper's leaf sits under 'Pathwise Gradient Estimation Methods,' distinct from model-based approaches that use learned dynamics models and from discrete action techniques that handle combinatorial spaces. By focusing on continuous actions and on-policy data, the work diverges from hybrid methods requiring replay buffers and from model-based rollouts, occupying a boundary between classical score-function estimators and fully differentiable simulation-based methods.

Among fourteen candidates examined, the first contribution (on-policy pathwise gradients without replay) showed no clear refutation across four candidates, suggesting relative novelty in this specific formulation. The second contribution (joint entropy and KL-constrained objective) examined ten candidates and found two refutable instances, indicating some overlap with prior regularization schemes. The third contribution (architectural components for value learning) was not directly assessed against prior work. The limited search scope—fourteen candidates from semantic search and citation expansion—means these findings reflect top-ranked matches rather than exhaustive coverage of the field.

Overall, the analysis suggests the work occupies a sparsely populated research direction, with the core algorithmic framework appearing relatively novel but the regularization objective showing partial overlap with existing methods. The small taxonomy leaf size and limited refutation evidence point toward a contribution that extends known ideas into a less-explored on-policy setting, though the restricted search scope leaves open the possibility of additional relevant prior work not captured in this analysis.

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

Contribution 1: On-policy algorithm using pathwise policy gradients without replay buffers

Description: The authors introduce REPPO, an on-policy reinforcement learning algorithm that learns state-action value functions from on-policy data alone, enabling the use of pathwise policy gradients without requiring large replay buffers typical of off-policy methods.

This contribution was assessed against **4 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Deep policy gradient methods without batch updates, target networks, or replay buffers

URL: View paper

Brief Assessment

Deep Policy Without Batch[1] focuses on incremental learning methods that eliminate batch updates, target networks, and replay buffers entirely. While both papers address on-policy learning without large replay buffers, the candidate's incremental approach (processing single samples) differs fundamentally from REPPO's on-policy batch approach (using rollout data with multiple epochs of updates).

---

2. Global Optimality Guarantees For Policy Gradient Methods

URL: View paper

Brief Assessment

Global Policy Gradient[7] focuses on theoretical convergence guarantees for policy gradient methods in general, not on the specific algorithmic innovation of combining on-policy learning with pathwise gradients to avoid replay buffers.

---

3. On-Policy Policy Gradient Reinforcement Learning Without On-Policy Sampling

URL: View paper

Brief Assessment

OnPolicy Without OnPolicy[9] focuses on adaptive off-policy sampling to reduce sampling error in on-policy data collection, not on learning state-action value functions from on-policy data alone to enable pathwise policy gradients. The candidate addresses a different technical problem (sampling distribution optimization) than the original contribution (value function learning for pathwise gradients).

---

4. Nash Policy Gradient: A Policy Gradient Method with Iteratively Refined Regularization for Finding Nash Equilibria

URL: View paper

Brief Assessment

Nash Policy Gradient[8] focuses on finding Nash equilibria in two-player zero-sum games through iterative regularization refinement, not on-policy RL with pathwise gradients for general control tasks.

Contribution 2: Joint entropy and KL-constrained policy optimization objective

Description: The authors develop a policy optimization framework that combines maximum entropy exploration with KL-divergence constraints on policy updates, automatically tuning both multipliers to balance exploration and stable learning.

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. Perception-aware policy optimization for multimodal reasoning

URL: View paper

Brief Assessment

Perception Aware Policy[15] focuses on multimodal reasoning with a KL divergence term for visual grounding (maximizing KL between original and corrupted visual inputs), not on combining maximum entropy exploration with KL-divergence constraints for general policy updates as in the original paper's contribution.

2. Equivalence between policy gradients and soft q-learning

URL: View paper

Prior Art Analysis

Policy Gradients Soft Q[18] demonstrates that entropy-regularized reinforcement learning with KL-divergence constraints was established prior to the original paper's work. The candidate paper presents a comprehensive framework combining maximum entropy objectives with KL-divergence penalties, showing both theoretical foundations and practical implementations. The paper derives the optimal policy under entropy regularization with KL constraints (Equation 1-6), defines entropy-augmented returns with KL penalties (Section 3), and demonstrates equivalence between soft Q-learning and policy gradient methods in this setting. This prior work establishes the mathematical framework and practical viability of combining entropy maximization with KL constraints for policy optimization.

Evidence

Evidence 1 - **Rationale**: The candidate establishes the entropy-regularized RL framework with KL penalties, citing prior work that predates the original paper, demonstrating this is not a novel contribution. - **Original**: reppo builds on the maximum entropy framework (ziebart et al., 2008) to encourage exploration. it combines this with a kl regularization scheme, inspired by the relative entropy policy search method (peters et al., 2010), which prevents aggressive policy updates from destabilizing the optimization. - **Candidate**: we shall consider an entropy-regularized version of the reinforcement learning problem, following various prior work (ziebart [2010], fox et al. [2015], haarnoja et al. [2017], nachum et al. [2017]). specifically, let us define the entropy-augmented return to be ∑∞ t=0 γt(rt -τklt) where rt is the rew...

Evidence 2 - **Rationale**: This demonstrates the candidate paper derives the relationship between entropy and KL terms in policy optimization, showing the mathematical framework existed before the original work. - **Original**: policy updates and multiplier tuning in the constrained objective, we introduce two hyperparameters, εh and εkl, which bound the entropy and kl divergence. the goal of the lagrangian parameters is to ensure that the policy stays close to these constraints. - **Candidate**: the boltzmann policy can be written as πb qθ(a) = π(a) exp((qθ(a) -vθ)/τ). (12) note that the term τlog ea∼π[exp(¯r(a)/τ)], appeared earlier in equation (6)). repeating the calculation from equation (2) through equation (6), but with qθ instead of ¯r, vθ = ea∼πbqθ [qθ(a)] -τdkl [πb qθ π]

3. A unified view of entropy-regularized Markov decision processes

URL: View paper

Prior Art Analysis

Unified Entropy Regularized[13] demonstrates that combining maximum entropy exploration with KL-divergence constraints in policy optimization was explored prior to the original paper's submission. The candidate paper presents a general framework for entropy-regularized reinforcement learning that explicitly combines entropy regularization with constrained policy updates, showing strong duality between entropy-regularized objectives and regularized Bellman equations. The candidate establishes theoretical foundations for jointly tuning entropy and KL parameters in policy optimization, which directly relates to the original paper's claimed contribution of developing a framework that 'combines maximum entropy exploration with KL-divergence constraints on policy updates, automatically tuning both multipliers.'

Evidence

Evidence 1 - **Rationale**: This shows that entropy-regularized policy optimization with automatic constraint adaptation was an established research direction before the original paper, contradicting the novelty claim of automatically tuning both multipliers. - **Original**: inspired by haarnoja et al. (2019), reppo automatically adapts these constraints when the policy violates them. - **Candidate**: the idea of entropy regularization has also been used extensively in the reinforcement learning literature (sutton and barto, 1998; szepesvári, 2010). entropyregularized variants of the classic bellman equations and the entailing reinforcementlearning algorithms have been proposed to induce safe exp...

4. Independent natural policy gradient methods for potential games: Finite-time global convergence with entropy regularization

URL: View paper

Brief Assessment

Independent Natural Potential[19] focuses on entropy-regularized natural policy gradient methods for potential games with fixed entropy and KL parameters (εh and εkl), not on automatically tuning both multipliers to balance exploration and stable learning as in the original paper's contribution.

5. The entropy mechanism of reinforcement learning for reasoning language models

URL: View paper

Brief Assessment

Entropy Reasoning Language[11] focuses on entropy dynamics and covariance-based regularization for LLM reasoning tasks, not on joint tuning of entropy and KL multipliers for general RL frameworks as in the original paper.

6. Fast Global Convergence of Natural Policy Gradient Methods with Entropy Regularization

URL: View paper

Brief Assessment

Fast Natural Policy[12] focuses on theoretical convergence guarantees for entropy-regularized natural policy gradient methods in tabular settings, not on the practical joint tuning of entropy and KL constraints for policy optimization that the original paper proposes.

7. Model-free deep reinforcement learningâ��algorithms and applications

URL: View paper

Brief Assessment

Model Free Deep RL[16] appears to discuss maximum entropy RL and KL constraints separately, but the provided context fragments are too limited to establish whether it presents the joint automatic tuning framework described in the original paper.

8. Proximal Policy Optimization with Entropy Regularization

URL: View paper

Brief Assessment

PPO Entropy Regularization[14] focuses solely on adding entropy regularization to PPO for exploration-exploitation balance, without incorporating KL-divergence constraints or automatic multiplier tuning that jointly balances both objectives.

9. Trust region policy optimization via entropy regularization for Kullback-Leibler divergence constraint

URL: View paper

Brief Assessment

Trust Region Entropy[10] focuses on regularizing TRPO's KL constraint via Shannon entropy to enhance exploration, whereas the original paper combines maximum entropy exploration with KL constraints using automatic multiplier tuning for both terms jointly.

10. Fast rates for maximum entropy exploration

URL: View paper

Brief Assessment

Fast Maximum Entropy[17] focuses on maximum entropy exploration in MDPs without KL-divergence constraints on policy updates. The paper addresses entropy maximization for exploration but does not combine it with KL-constrained policy optimization as described in the original contribution.

Contribution 3: Evaluation of architectural components for on-policy value learning

Description: The authors assess the impact of recent neural network design advances including categorical Q-learning with cross-entropy losses, normalized architectures, and auxiliary tasks on stabilizing value function learning in the on-policy setting.

This contribution was assessed against **0 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

Appendix: Text Similarity Detection

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

References

- [0] Relative Entropy Pathwise Policy Optimization View paper
- [1] Deep policy gradient methods without batch updates, target networks, or replay buffers View paper
- [2] Differentiable discrete event simulation for queuing network control View paper
- [3] PIPPS: Flexible Model-Based Policy Search Robust to the Curse of Chaos View paper
- [4] A Differential and Pointwise Control Approach to Reinforcement Learning View paper
- [5] Deterministic Policy Optimization by Combining Pathwise and Score Function Estimators for Discrete Action Spaces View paper
- [6] Control-based theories for credit assignment in neural networks and reinforcement learning View paper
- [7] Global Optimality Guarantees For Policy Gradient Methods View paper
- [8] Nash Policy Gradient: A Policy Gradient Method with Iteratively Refined Regularization for Finding Nash Equilibria View paper
- [9] On-Policy Policy Gradient Reinforcement Learning Without On-Policy Sampling View paper
- [10] Trust region policy optimization via entropy regularization for Kullback-Leibler divergence constraint View paper
- [11] The entropy mechanism of reinforcement learning for reasoning language models View paper
- [12] Fast Global Convergence of Natural Policy Gradient Methods with Entropy Regularization View paper
- [13] A unified view of entropy-regularized Markov decision processes View paper
- [14] Proximal Policy Optimization with Entropy Regularization View paper
- [15] Perception-aware policy optimization for multimodal reasoning View paper
- [16] Model-free deep reinforcement learningâ□□algorithms and applications View paper
- [17] Fast rates for maximum entropy exploration View paper
- [18] Equivalence between policy gradients and soft q-learning View paper
- [19] Independent natural policy gradient methods for potential games: Finite-time global convergence with entropy regularization View paper