Novelty Assessment Report

Paper: Solving General-Utility Markov Decision Processes in the Single-Trial Regime with Online Planning

PDF URL: https://openreview.net/pdf?id=2XSP20jV0T

Venue: ICLR 2026 Conference Submission

Year: 2026

Report Generated: 2026-01-05

Abstract

In this work, we contribute the first approach to solve infinite-horizon discounted general-utility Markov decision processes (GUMDPs) in the single-trial regime, i.e., when the agent's performance is evaluated based on a single trajectory. First, we provide some fundamental results regarding policy optimization in the single-trial regime, investigating which class of policies suffices for optimality, casting our problem as a particular MDP that is equivalent to our original problem, as well as studying the computational hardness of policy optimization in the single-trial regime. Second, we show how we can leverage online planning techniques, in particular a Monte-Carlo tree search algorithm, to solve GUMDPs in the single-trial regime. Third, we provide experimental results showcasing the superior performance of our approach in comparison to relevant baselines.

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: **Single-Trial Policy Optimization in General-Utility Markov Decision Processes**

A total of **18 papers** were analyzed and organized into a taxonomy with **17 categories**.

Taxonomy Overview

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

- **Single-Trajectory Learning Algorithms**
- **General-Utility and Constrained MDP Optimization**
- **Policy Gradient and Multi-Agent Methods**
- **Representation and Structure Learning in MDPs**
- **Navigation and Optimal Policy Identification**
- **Robotic Manipulation Applications**
- **Foundational RL Concepts and Sequential Learning**

Complete Taxonomy Tree

- Single-Trial Policy Optimization in General-Utility Markov Decision Processes Survey Taxonomy
- Single-Trajectory Learning Algorithms
  - Value-Function-Based Single-Trajectory Methods (3 papers)
  - [8] An incremental off-policy search in a model-free Markov decision process using a single sample path (Ajin George Joseph, 2018) View paper
  - [10] Learning near-optimal policies with Bellman-residual minimization based fitted policy iteration and a single sample path (A. Antos, 2008) View paper
  - [16] Value-iteration based fitted policy iteration: learning with a single trajectory (AndrÃ¡s Antos, 2007) View paper
  - Distributionally Robust Single-Trajectory RL (1 papers)
  - [4] Single-Trajectory Distributionally Robust Reinforcement Learning (Liang Zhipeng, 2023) View paper
  - Learning-Augmented Single-Trajectory Methods (1 papers)
  - [15] Beyond Black-Box Advice: Learning-Augmented Algorithms for MDPs with Q-Value Predictions (Li Tongxin, 2023) View paper
  - Theoretical Foundations and Complexity Bounds (1 papers)
  - [18] Learning from A Single Markovian Trajectory: Optimality and Variance Reduction (Z Sun, n.d.) View paper
- General-Utility and Constrained MDP Optimization
  - Single-Trial General-Utility MDP Planning ★ (1 papers)
  - [0] Solving General-Utility Markov Decision Processes in the Single-Trial Regime with Online Planning (Anon et al., 2026) View paper
  - Constrained MDP Policy Optimization (1 papers)
  - [2] Finite-Time Complexity of Online Primal-Dual Natural Actor-Critic Algorithm for Constrained Markov Decision Processes (Sihan Zeng, 2021) View paper
  - Occupancy-Based and Scalable General-Utility Methods (1 papers)
  - [5] Towards Scalable General Utility Reinforcement Learning: Occupancy Approximation, Sample Complexity and Global Optimality (A Barakat, 2024) View paper
  - Behavioral Alignment via Prospect-Theoretic Utilities (1 papers)
  - [12] A Prospect-Theoretic Policy Gradient Algorithm for Behavioral Alignment in Reinforcement Learning (O Lepel, 2024) View paper
- Policy Gradient and Multi-Agent Methods (1 papers)
  - [7] MDPGT: Momentum-based Decentralized Policy Gradient Tracking (Balu, 2022) View paper
- Representation and Structure Learning in MDPs
  - Exogenous Block MDP Representation Learning (1 papers)

    - [13] Learning a Fast Mixing Exogenous Block MDP using a Single Trajectory (Levine, 2024) View paper
    - Task-Agnostic Exploration via Entropy Maximization (1 papers)
    - [17] Task-agnostic exploration via maximum state entropy policy optimization (Pratissoli, 2019) View paper
- Navigation and Optimal Policy Identification (1 papers)
    - [1] Navigating to the best policy in markov decision processes (Aymen Al Marjani, 2021) View paper
- Robotic Manipulation Applications
    - Single-Demonstration Dexterous Grasping (1 papers)
    - [3] DemoGrasp: Universal Dexterous Grasping from a Single Demonstration (YUAN HaoQi, 2025) View paper
    - Guided Multi-Contact Loco-Manipulation (1 papers)
    - [6] Guided Reinforcement Learning for Robust Multi-Contact Loco-Manipulation (Sleiman, 2024) View paper
    - Insertion Skill Learning from One-Shot Demonstration (1 papers)
    - [9] Skill learning for robotic insertion based on one-shot demonstration and reinforcement learning (Ying Li, 2021) View paper
    - Process Reward Modeling for High-Precision Manipulation (1 papers)
    - [11] Robo-Dopamine: General Process Reward Modeling for High-Precision Robotic Manipulation (Huajie Tan, 2025) View paper
- Foundational RL Concepts and Sequential Learning (1 papers)
    - [14] 2 reinforcement learning and its (GB ANDREW, 2004) View paper

Narrative

Core task: single-trial policy optimization in general-utility Markov decision processes. The field addresses how to learn effective policies when only a single trajectory is available and when the objective extends beyond standard expected cumulative reward. The taxonomy reveals several complementary perspectives: Single-Trajectory Learning Algorithms focus on data-efficient methods that extract maximal information from one rollout, while General-Utility and Constrained MDP Optimization tackles non-standard objectives such as risk-sensitive criteria, constraints, or prospect-theoretic utilities. Policy Gradient and Multi-Agent Methods explore gradient-based and distributed optimization, Representation and Structure Learning in MDPs examines how to discover useful state abstractions or exploit problem structure, and Navigation and Optimal Policy Identification as well as Robotic Manipulation Applications ground these ideas in concrete domains. Foundational RL Concepts and Sequential Learning provides the theoretical bedrock, connecting classical results to modern challenges in sample-starved or non-standard settings.

Within this landscape, a particularly active line of work investigates how to handle general utility functions and constraints when data is scarce. For instance, Scalable General Utility RL[5] and Prospect Theoretic Policy Gradient[12] illustrate recent efforts to optimize beyond expected return, while Single Trajectory Robust RL[4] and Single Markovian Trajectory[18] emphasize robustness and sample efficiency from minimal data. General Utility Online Planning[0] sits squarely in this intersection, addressing the challenge of planning under general utilities when only one trial is available. Its emphasis on online planning distinguishes it from offline or batch approaches such as Primal Dual Actor Critic[2], which typically assume richer data or iterative policy updates. By targeting single-trial scenarios with non-standard objectives, General Utility Online Planning[0] complements works like Scalable General Utility RL[5] that scale to larger problems but may rely on more extensive sampling, highlighting an ongoing trade-off between sample complexity and the breadth of utility functions one can handle.

Related Works in Same Category

No sibling papers were found in the same taxonomy leaf. A taxonomy-subtopic-level comparison will be produced instead.

Taxonomy-Level Summary

The original leaf focuses on online planning methods for single-trajectory evaluation in infinite-horizon discounted general-utility MDPs, distinguishing itself from multi-trial approaches. The sibling subtopics cover complementary aspects of general-utility MDPs: behavioral alignment using prospect theory, constrained optimization with utility constraints, and occupancy-based methods with sample complexity guarantees. Together, these categories partition the general-utility MDP space by evaluation paradigm (single vs. multi-trial), constraint structure (unconstrained vs. constrained), and utility specification (behavioral economics vs. standard general utilities).

Similarities: - All categories address MDPs with general utility functions beyond standard cumulative reward maximization - All involve infinite-horizon discounted settings with policy optimization objectives - Each category explicitly excludes standard reward MDPs, indicating focus on richer utility structures - All represent algorithmic approaches to decision-making under utility-based objectives

Differences: - Single-Trial General-Utility MDP Planning evaluates policies on single trajectories with online planning, while Occupancy-Based methods use multi-trial sample complexity guarantees and occupancy measure approximations - Constrained MDP Policy Optimization handles utility constraints via primal-dual methods, whereas the original leaf addresses unconstrained general-utility optimization - Behavioral Alignment incorporates prospect theory and human-aligned utilities from behavioral economics, while the original leaf uses general utility functions without specific behavioral modeling - The original leaf emphasizes online planning methods, while Occupancy-Based approaches focus on scalable convex optimization and policy gradients with theoretical guarantees

Suggested Search Directions: - Hybrid methods combining single-trial planning with occupancy-based approximations for improved sample efficiency - Extensions of single-trial planning to handle soft constraints or behavioral utilities - Theoretical connections between online planning regret bounds and sample complexity in general-utility settings

Sibling Subtopics

- **Behavioral Alignment via Prospect-Theoretic Utilities** (leaves: 1, papers: 1)
- Scope: Policy gradient algorithms incorporating prospect theory or behavioral economics utilities for human-aligned decision-making.
- Exclude: Standard utility functions belong under Occupancy-Based or Constrained methods; single-trial planning belongs under Single-Trial General-Utility MDP Planning.
- **Constrained MDP Policy Optimization** (leaves: 1, papers: 1)
- Scope: Algorithms optimizing discounted reward subject to constraints on discounted cumulative utilities using primal-dual or actor-critic methods.
- Exclude: Unconstrained general-utility MDPs belong under Single-Trial General-Utility MDP Planning; standard reward maximization belongs elsewhere.
- **Occupancy-Based and Scalable General-Utility Methods** (leaves: 1, papers: 1)
- Scope: Approaches using occupancy measure approximations or policy gradients for scalable convex general-utility RL with sample complexity guarantees.
- Exclude: Single-trial planning methods belong under Single-Trial General-Utility MDP Planning; constrained methods belong under Constrained MDP Policy Optimization.

Contributions Analysis

Overall novelty summary. ```json { "paragraphs": ["The paper introduces an approach for solving infinite-horizon discounted general-utility MDPs evaluated on single trajectories, combining non-standard utility functions with stringent data constraints. According to the taxonomy tree, this work occupies the 'Single-Trial General-Utility MDP Planning' leaf under 'General-Utility and Constrained MDP Optimization'. Notably, this leaf contains only the original paper itself—no sibling papers appear in the same node—indicating that this precise intersection of single-trial evaluation and general-utility objectives represents a sparse research direction within the surveyed literature.",

"The taxonomy reveals related but distinct directions nearby: 'Constrained MDP Policy Optimization' addresses utility constraints via primal-dual methods but typically assumes multi-trial settings, while 'Occupancy-Based and Scalable General-Utility Methods' handle general utilities with sample complexity guarantees yet do not focus on single-trial constraints. Under 'Single-Trajectory Learning Algorithms', sibling leaves such as 'Value-Function-Based Single-Trajectory Methods' and 'Distributionally Robust Single-Trajectory RL' tackle data scarcity but assume standard reward objectives. The paper bridges these communities by merging general-utility optimization with single-trajectory planning, a combination not directly covered by existing taxonomy nodes.",

"Among the three contributions analyzed from 24 candidate papers, 'Fundamental results for policy optimization in the single-trial regime' examined 4 candidates and found 1 potentially refutable overlap, suggesting that basic theoretical characterizations may have some prior coverage. In contrast, 'Monte-Carlo tree search algorithm for single-trial GUMDPs' and 'First approach to solve infinite-horizon discounted GUMDPs in single-trial regime' each examined 10 candidates with zero refutable overlaps, indicating that the algorithmic and problem-formulation aspects appear more novel within this limited search scope. The statistics suggest the core methodological and application-level claims face less direct competition in the examined literature.",

"Given the constrained search scale—24 candidates from top-K semantic matches and citation expansion—this analysis captures the immediate neighborhood rather than an exhaustive survey. The absence of sibling papers in the same taxonomy leaf and the low refutation counts for algorithmic contributions suggest the work occupies a relatively unexplored niche at the intersection of single-trial constraints and general utilities. However, broader searches or domain-specific venues might reveal additional relevant work not surfaced here."] } ```

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

Contribution 1: Fundamental results for policy optimization in the single-trial regime

Description: The authors establish theoretical foundations for solving GUMDPs in the single-trial setting. They prove that non-Markovian policies are necessary for optimality, introduce an occupancy MDP formulation that is equivalent to the original problem, and demonstrate that policy optimization in this regime is NP-Hard even for smooth convex objectives.

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. Reinforcement learning for stochastic shortest path with dead-ends

URL: View paper

Brief Assessment

Shortest Path Dead Ends[22] focuses on stochastic shortest path problems with dead-end states in goal-oriented tasks, not general-utility MDPs in single-trial regimes. The technical frameworks differ fundamentally.

---

2. Stopping the Revolving Door: MDP-Based Decision Support for Community Corrections Placement

URL: View paper

Brief Assessment

Community Corrections MDP[20] focuses on community corrections placement decisions using standard MDPs, not general-utility MDPs in single-trial regimes. The candidate does not address non-Markovian policies, occupancy MDP formulations, or computational hardness results for single-trial GUMDPs.

---

3. The importance of non-markovianity in maximum state entropy exploration

URL: View paper

Prior Art Analysis

Non Markovianity Entropy Exploration[19] demonstrates that similar theoretical foundations for single-trial policy optimization were established prior to the original paper. Both papers prove that non-Markovian policies are necessary for optimality in single-trial settings, establish equivalence through reformulated MDPs (extended MDP vs occupancy MDP), and prove NP-hardness of the optimization problem. The candidate paper explicitly addresses the single-trial regime with finite-sample objectives, proves non-Markovian policy necessity, introduces an extended MDP formulation equivalent to the original problem, and establishes computational hardness - all core claims of the original paper's contribution.

Evidence

Evidence 1 - **Rationale**: Both papers prove that non-Markovian policies strictly dominate Markovian policies in the single-trial setting, establishing the necessity of non-Markovianity for optimal performance. - **Original**: theorem 1.there exists a gumdpm f withγ∈(0,1)andl-lipschitz convex objective such that (lower is better): 1.f 1(πs)> f1(πm), for someπ m ∈π m and anyπ s ∈π s. 2.f 1(πm)> f1(πnm), for someπ nm ∈π nm and anyπ m ∈π m. The result above shows that, in general, the class of stationary policies is strictly... - **Candidate**: theorem 4.2 (non-markovian optimality). for every cmp mand trajectory ht ∈h[t], there exists a deterministic non-markovian policy πnm ∈πd nm that suffers zero regretto-go rt-t(πnm,ht) = 0, whereas for any πm ∈πm we have rt-t(πm,ht) ≥0. corollary 4.3 (sufficient condition). for every cmp m and traject...

Evidence 2 - **Rationale**: Both papers establish NP-hardness of finding optimal non-Markovian policies in the single-trial regime, proving computational intractability of the policy optimization problem. - **Original**: theorem 3(np-hardness of policy optimization in the single-trial regime).given a gumdp with objectivef 1,h and a threshold valueλ∈r, it is np-hard to determine whether there exists a policy π∈π d nm satisfyingf 1,h(π)≤λ. - **Candidate**: theorem 5.4. ψ0 is np-hard. proof sketch. to prove the theorem, it is sufficient to show that there exists a problemψc ∈np-hard so thatψc ≤p ψ0. we show this by reducing 3sat, which is a well-known np-complete problem, to ψ0.

Evidence 3 - **Rationale**: Both papers reformulate the objective to focus on single-trial empirical occupancies/visitation frequencies rather than expected occupancies over infinite trials, establishing the same conceptual foundation. - **Original**: we consider the setting in which the agent interacts with its environment over a single-trial, i.e., a single trajectory. for a given policyπ∈π nm, we introduce the random vectord π : ω→Δ(s x a), which corresponds to the empirical discounted state-action occupancy associated with the probability spa... - **Candidate**: in this section, we reformulate the typical maximum state entropy exploration objective of a cmp (1) to account for a finite-sample regime. crucially, we consider the expected entropy of the state visitation frequency rather than the entropy of the expected state visitation frequency, which results i...

---

4. Convex reinforcement learning in finite trials

URL: View paper

Brief Assessment

[Final Audit Failure] The model insisted on a refutation claim but failed to provide verifiable evidence after multiple retries. Marked as cannot_refute for safety. Please manually verify the candidate text.

Contribution 2: Monte-Carlo tree search algorithm for single-trial GUMDPs

Description: The authors develop an MCTS-based online planning algorithm that solves the occupancy MDP formulation. The algorithm provably converges to the optimal action at each timestep given sufficient iterations and achieves polynomial regret bounds.

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 Bayesian approach to online planning

URL: View paper

Brief Assessment

Bayesian Online Planning[24] focuses on Bayesian approaches to MCTS with uncertainty quantification for neural network predictions, not on solving general-utility MDPs in the single-trial regime with occupancy formulations.

2. Dec-MCTS: Decentralized planning for multi-robot active perception

URL: View paper

Brief Assessment

Decentralized MCTS[32] focuses on multi-robot active perception with decentralized planning and intermittent communication, not single-trial general-utility MDPs with occupancy tracking and non-Markovian policies.

3. Monte-Carlo tree search and rapid action value estimation in computer Go

URL: View paper

Brief Assessment

MCTS Rapid Action Value[26] focuses on computer Go with rapid action value estimation techniques, not on general-utility MDPs or single-trial regime optimization with occupancy tracking.

4. Bayesian Inference in Monte-Carlo Tree Search

URL: View paper

Brief Assessment

Bayesian MCTS[30] focuses on Bayesian inference for bandit-tree problems with binary rewards, not general-utility MDPs in single-trial regimes. The candidate addresses uncertainty estimation in MCTS through posterior distributions, while the original develops MCTS for occupancy MDPs with general utility functions.

5. Monte carlo tree search in continuous spaces using voronoi optimistic optimization with regret bounds

URL: View paper

Brief Assessment

MCTS Continuous Voronoi[29] addresses continuous action spaces in deterministic planning with MCTS using Voronoi optimization, not the single-trial GUMDP regime with occupancy tracking that the original paper focuses on.

6. Lipschitz Lifelong Monte Carlo Tree Search for Mastering Non-Stationary Tasks

URL: View paper

Brief Assessment

Lipschitz Lifelong MCTS[31] addresses lifelong planning across non-stationary task sequences, not single-trial GUMDPs with occupancy-based objectives. The candidate focuses on knowledge transfer between tasks with varying dynamics, while the original develops MCTS for general-utility objectives evaluated on single trajectories.

7. Convergence of Monte Carlo Tree Search in Simultaneous Move Games

URL: View paper

Brief Assessment

MCTS Simultaneous Move Games[27] focuses on convergence guarantees for MCTS in zero-sum games with simultaneous moves, not on solving general-utility MDPs in the single-trial regime with occupancy-based objectives.

8. Pareto monte carlo tree search for multi-objective informative planning

URL: View paper

Brief Assessment

Pareto MCTS[25] addresses multi-objective informative planning for environmental monitoring, not general-utility MDPs in single-trial regimes. The candidate focuses on Pareto optimization across competing objectives (exploration vs exploitation) rather than solving GUMDPs with convergence guarantees and regret bounds for single-trial evaluation.

9. Monte-Carlo tree search with uncertainty propagation via optimal transport

URL: View paper

Brief Assessment

MCTS Uncertainty Propagation[23] focuses on propagating uncertainty via Wasserstein barycenters in stochastic/partially observable MDPs, not on solving general-utility MDPs in the single-trial regime with occupancy-based formulations.

10. Monte-Carlo tree search for constrained POMDPs

URL: View paper

Brief Assessment

MCTS Constrained POMDPs[28] addresses constrained POMDPs with multiple cost constraints, not general-utility MDPs in the single-trial regime. The candidate focuses on multi-objective optimization with cost constraints rather than general utility functions over occupancy measures.

Contribution 3: First approach to solve infinite-horizon discounted GUMDPs in single-trial regime

Description: The authors present the first method for solving infinite-horizon discounted GUMDPs when performance is evaluated on a single trajectory rather than over infinite trials. This addresses a practical limitation where optimal policies for the infinite-trial formulation may perform poorly when evaluated on limited trajectories.

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. Navigating to the best policy in markov decision processes

URL: View paper

Brief Assessment

Navigating Best Policy[1] addresses a different problem: active pure exploration in standard MDPs to identify the best policy, not solving GUMDPs with general utility functions in the single-trial regime. The candidate focuses on sample complexity for policy identification in communicating MDPs, while the original work tackles general-utility objectives evaluated on single trajectories.

2. Experimenting on markov decision processes with local treatments

URL: View paper

Brief Assessment

Local Treatment MDPs[37] focuses on A/B testing in standard MDPs with local treatments, not on general-utility MDPs (GUMDPs) where objectives are non-linear functions of occupancy measures. The candidate addresses treatment locality in standard reward settings, while the original tackles a fundamentally different problem class.

3. Policy gradient for continuing tasks in discounted markov decision processes

URL: View paper

Brief Assessment

Policy Gradient Continuing Tasks[33] addresses standard infinite-horizon discounted MDPs with policy gradient methods for continuing tasks without restarts, not the GUMDP framework with general utility functions evaluated on single trajectories.

4. Last-Iterate Convergent Policy Gradient Primal-Dual Methods for Constrained MDPs

URL: View paper

Brief Assessment

Last Iterate Primal Dual[40] addresses constrained MDPs with constraint satisfaction, not general-utility MDPs evaluated on single trajectories. The frameworks differ fundamentally in their problem formulations and objectives.

5. An online actorâ□□critic algorithm with function approximation for constrained markov decision processes

URL: View paper

Brief Assessment

Online Actor Critic Constrained[34] addresses constrained MDPs with function approximation in online settings, but does not specifically target GUMDPs or the single-trial evaluation regime that is central to the original paper's contribution.

6. On the Convergence of Single-Timescale Actor-Critic

URL: View paper

Brief Assessment

Single Timescale Actor Critic[38] focuses on standard infinite-horizon discounted MDPs with linear reward functions, not general-utility MDPs (GUMDPs) where objectives are non-linear functions of occupancy measures evaluated on single trajectories.

7. On One Adjoint Trajectory in Infinite-Horizon Control Problems

URL: View paper

Brief Assessment

Adjoint Trajectory Infinite Horizon[41] focuses on optimal control problems with adjoint trajectories in infinite-horizon settings, not on general-utility MDPs or single-trajectory evaluation regimes. No full text was provided to assess overlap.

8. Efficiently breaking the curse of horizon in off-policy evaluation with double reinforcement learning

URL: View paper

Brief Assessment

Breaking Horizon Curse[36] focuses on off-policy evaluation in standard MDPs with time-invariant structure, not on solving general-utility MDPs (GUMDPs) where objectives are non-linear functions of occupancy measures in single-trial settings.

9. Finite-Time Analysis of Asynchronous Stochastic Approximation and -Learning

URL: View paper

Brief Assessment

Asynchronous Stochastic Approximation[39] focuses on finite-time convergence analysis of asynchronous Q-learning for standard MDPs with discounted rewards, not on general-utility MDPs (GUMDPs) where objectives are non-linear functions of occupancy measures evaluated on single trajectories.

10. Stochastic optimization methods for policy evaluation in reinforcement learning

URL: View paper

Brief Assessment

Stochastic Policy Evaluation[35] focuses on policy evaluation methods for standard MDPs using stochastic optimization techniques (TD learning, gradient TD, variance reduction). It does not address GUMDPs, single-trial vs. infinite-trial formulations, or the specific problem of optimizing policies when performance depends on empirical occupancies from single trajectories.

Appendix: Text Similarity Detection

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

References

- [0] Solving General-Utility Markov Decision Processes in the Single-Trial Regime with Online Planning View paper
- [1] Navigating to the best policy in markov decision processes View paper
- [2] Finite-Time Complexity of Online Primal-Dual Natural Actor-Critic Algorithm for Constrained Markov Decision Processes View paper
- [3] DemoGrasp: Universal Dexterous Grasping from a Single Demonstration View paper
- [4] Single-Trajectory Distributionally Robust Reinforcement Learning View paper
- [5] Towards Scalable General Utility Reinforcement Learning: Occupancy Approximation, Sample Complexity and Global Optimality View paper

- [6] Guided Reinforcement Learning for Robust Multi-Contact Loco-Manipulation View paper
- [7] MDPGT: Momentum-based Decentralized Policy Gradient Tracking View paper
- [8] An incremental off-policy search in a model-free Markov decision process using a single sample path View paper
- [9] Skill learning for robotic insertion based on one-shot demonstration and reinforcement learning View paper
- [10] Learning near-optimal policies with Bellman-residual minimization based fitted policy iteration and a single sample path View paper
- [11] Robo-Dopamine: General Process Reward Modeling for High-Precision Robotic Manipulation View paper
- [12] A Prospect-Theoretic Policy Gradient Algorithm for Behavioral Alignment in Reinforcement Learning View paper
- [13] Learning a Fast Mixing Exogenous Block MDP using a Single Trajectory View paper
- [14] 2 reinforcement learning and its View paper
- [15] Beyond Black-Box Advice: Learning-Augmented Algorithms for MDPs with Q-Value Predictions View paper
- [16] Value-iteration based fitted policy iteration: learning with a single trajectory View paper
- [17] Task-agnostic exploration via maximum state entropy policy optimization View paper
- [18] Learning from A Single Markovian Trajectory: Optimality and Variance Reduction View paper
- [19] The importance of non-markovianity in maximum state entropy exploration View paper
- [20] Stopping the Revolving Door: MDP-Based Decision Support for Community Corrections Placement View paper
- [21] Convex reinforcement learning in finite trials View paper
- [22] Reinforcement learning for stochastic shortest path with dead-ends View paper
- [23] Monte-Carlo tree search with uncertainty propagation via optimal transport View paper
- [24] A Bayesian approach to online planning View paper
- [25] Pareto monte carlo tree search for multi-objective informative planning View paper
- [26] Monte-Carlo tree search and rapid action value estimation in computer Go View paper
- [27] Convergence of Monte Carlo Tree Search in Simultaneous Move Games View paper
- [28] Monte-Carlo tree search for constrained POMDPs View paper
- [29] Monte carlo tree search in continuous spaces using voronoi optimistic optimization with regret bounds View paper
- [30] Bayesian Inference in Monte-Carlo Tree Search View paper
- [31] Lipschitz Lifelong Monte Carlo Tree Search for Mastering Non-Stationary Tasks View paper
- [32] Dec-MCTS: Decentralized planning for multi-robot active perception View paper
- [33] Policy gradient for continuing tasks in discounted markov decision processes View paper
- [34] An online actorâ□□critic algorithm with function approximation for constrained markov decision processes View paper
- [35] Stochastic optimization methods for policy evaluation in reinforcement learning View paper
- [36] Efficiently breaking the curse of horizon in off-policy evaluation with double reinforcement learning View paper
- [37] Experimenting on markov decision processes with local treatments View paper
- [38] On the Convergence of Single-Timescale Actor-Critic View paper
- [39] Finite-Time Analysis of Asynchronous Stochastic Approximation and -Learning View paper
- [40] Last-Iterate Convergent Policy Gradient Primal-Dual Methods for Constrained MDPs View paper
- [41] On One Adjoint Trajectory in Infinite-Horizon Control Problems View paper