Novelty Assessment Report

Paper: Q-Learning with Fine-Grained Gap-Dependent Regret
PDF URL: https://openreview.net/pdf?id=fE0RJto3Na
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2026-01-04

Abstract

We study fine-grained gap-dependent regret bounds for model-free reinforcement learning in episodic tabular Markov Decision Processes. Existing model-free algorithms achieve minimax worst-case regret, but their gap-dependent bounds remain coarse and fail to fully capture the structure of suboptimality gaps. We address this limitation by establishing fine-grained gap-dependent regret bounds for both UCB-based and non-UCB-based algorithms. In the UCB-based setting, we develop a novel analytical framework that explicitly separates the analysis of optimal and suboptimal state-action pairs, yielding the first fine-grained regret upper bound for UCB-Hoeffding (Jin et al., 2018). To highlight the generality of this framework, we introduce ULCB-Hoeffding, a new UCB-based algorithm inspired by AMB (Xu et al., 2021) but with a simplified structure, which enjoys fine-grained regret guarantees and empirically outperforms AMB. In the non-UCB-based setting, we revisit the only known algorithm AMB, and identify two key issues in its algorithm design and analysis: improper truncation in the Q-updates and violation of the martingale difference condition in its concentration argument. We propose a refined version of AMB that addresses these issues, establishing the first rigorous fine-grained gap-dependent regret for a non-UCB-based method, with experiments demonstrating improved performance over AMB.

Core Task Landscape

This paper addresses: **Fine-Grained Gap-Dependent Regret Bounds for Model-Free Reinforcement Learning**

A total of **37 papers** were analyzed and organized into a taxonomy with **19 categories**.

Taxonomy Overview

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

- **Tabular Episodic MDP Algorithms with Gap-Dependent Bounds**
- **Function Approximation and Structured MDPs**
- **Specialized Settings and Extensions**
- **Multi-Agent and Federated Reinforcement Learning**
- **General Frameworks and Theoretical Foundations**
- **Posterior Sampling and Bayesian Methods**
- **Discounted Infinite-Horizon and Continuous Settings**
- **Exploration Strategies and Practice-Based Learning**
- **Extensive-Form Games and Sequential Decision Making**
- **Robust and Adversarial Decision Making**
- ... and 1 more categories

Complete Taxonomy Tree

- Fine-Grained Gap-Dependent Regret Bounds for Model-Free Reinforcement Learning Survey Taxonomy
- Tabular Episodic MDP Algorithms with Gap-Dependent Bounds
  - UCB-Based Optimistic Algorithms ★ (4 papers)
  - [0] Q-Learning with Fine-Grained Gap-Dependent Regret (Anon et al., 2026) View paper
  - [7] Gap-dependent bounds for q-learning using reference-advantage decomposition (Zheng, 2024) View paper
  - [25] Tail Distribution of Regret in Optimistic Reinforcement Learning (Sajad Khodadadian, 2025) View paper
  - [28] Non-Asymptotic Gap-Dependent Regret Bounds for Tabular MDPs (Simchowitz, 2019) View paper
  - Bootstrap and Multi-Step Methods (1 papers)
  - [5] Fine-grained gap-dependent bounds for tabular mdps via adaptive multi-step bootstrap (Haike Xu, 2021) View paper
  - Non-UCB Optimistic Approaches (1 papers)
  - [19] Beyond Value-Function Gaps: Improved Instance-Dependent Regret Bounds for Episodic Reinforcement Learning (Dann, 2021) View paper
- Function Approximation and Structured MDPs
  - Linear Function Approximation (2 papers)
  - [18] Logarithmic regret for reinforcement learning with linear function approximation (He, 2021) View paper
  - [21] The best of both worlds: Reinforcement learning with logarithmic regret and policy switches (Velegkas, 2022) View paper
  - Nonlinear Function Approximation (2 papers)
  - [12] VOL: Towards Optimal Regret in Model-free RL with Nonlinear Function Approximation (A Agarwal, 2023) View paper
  - [14] Randomized exploration for reinforcement learning with multinomial logistic function approximation (Wooseong Cho, 2024) View paper
  - Hybrid and Model-Augmented Methods (2 papers)
  - [15] Hybrid reinforcement learning breaks sample size barriers in linear mdps (Wei Fan, 2024) View paper
  - [23] Can Q-learning be improved with advice? (Noah Golowich, 2022) View paper

- Specialized Settings and Extensions
  - Risk-Sensitive and Variance-Dependent Objectives (3 papers)
  - [1] A tighter problem-dependent regret bound for risk-sensitive reinforcement learning (X Hu, 2023) View paper
  - [10] Sharp Variance-Dependent Bounds in Reinforcement Learning: Best of Both Worlds in Stochastic and Deterministic Environments (Zhou, 2023) View paper
  - [17] Cascaded Gaps: Towards Gap-Dependent Regret for Risk-Sensitive Reinforcement Learning (Fei Yingjie, 2022) View paper
  - Non-Stationary and Adaptive Environments (4 papers)
  - [8] Efficient Restarts in Non-Stationary Model-Free Reinforcement Learning (Nonaka Hiroshi, 2025) View paper
  - [22] Model-Free Nonstationary Reinforcement Learning: Near-Optimal Regret and Applications in Multiagent Reinforcement Learning and Inventory Control (Weichao Mao, 2024) View paper
  - [29] Near-Optimal Regret Bounds for Model-Free RL in Non-Stationary Episodic MDPs (Weichao Mao, 2021) View paper
  - Offline and Pessimistic Learning (1 papers)
  - [2] Information-directed pessimism for offline reinforcement learning (A Koppel, 2024) View paper
- Multi-Agent and Federated Reinforcement Learning (4 papers)
  - [3] Gap-Dependent Bounds for Federated -learning (H Zhang, 2025) View paper
  - [4] Regret-Optimal Q-Learning with Low Cost for Single-Agent and Federated Reinforcement Learning (Zhang Haochen, 2025) View paper
  - [9] Gap-Dependent Regret for Federated Q-Learning (Zhang, 2025) View paper
  - [26] Advances in Inverse Problems and Decision-making under Uncertainty (Zheng, 2025) View paper
- General Frameworks and Theoretical Foundations
  - Instance-Dependent Complexity and Optimality (2 papers)
  - [11] Instance-optimality in interactive decision making: Toward a non-asymptotic theory (Wagenmaker, 2023) View paper
  - [32] Instance-Dependent Complexity of Contextual Bandits and Reinforcement Learning: A Disagreement-Based Perspective (Dylan J. Foster, 2020) View paper
  - Unified Algorithmic Frameworks (2 papers)
  - [6] Unified algorithms for RL with Decision-Estimation Coefficients: PAC, reward-free, preference-based learning and beyond (Fan Chen, 2025) View paper
  - [13] Politex: Regret bounds for policy iteration using expert prediction (Yasin Abbasi Yadkori, 2019) View paper
  - Model Selection and Regret Balancing (1 papers)
  - [34] Regret Bound Balancing and Elimination for Model Selection in Bandits and RL (Pacchiano, 2022) View paper
- Posterior Sampling and Bayesian Methods (1 papers)
  - [27] A Provably Efficient Model-Free Posterior Sampling Method for Episodic Reinforcement Learning (Dann, 2022) View paper
- Discounted Infinite-Horizon and Continuous Settings (3 papers)
  - [16] Adaptive discretization in online reinforcement learning (Sinclair, 2023) View paper
  - [30] Regret-Optimal Model-Free Reinforcement Learning for Discounted MDPs with Short Burn-In Time (Ji Xiang, 2023) View paper
  - [35] Online Regret Bounds for Undiscounted Continuous Reinforcement Learning (Ortner, 2013) View paper
- Exploration Strategies and Practice-Based Learning (1 papers)
  - [36] Efficient Reinforcement Learning via Initial Pure Exploration (Putta, 2017) View paper
- Extensive-Form Games and Sequential Decision Making (2 papers)
  - [24] Model-free online learning in unknown sequential decision making problems and games (Gabriele Farina, 2021) View paper
  - [31] Generalized Bandit Regret Minimizer Framework in Imperfect Information Extensive-Form Game (Meng, 2022) View paper
- Robust and Adversarial Decision Making (1 papers)
  - [20] Regret Bounds for Robust Online Decision Making (Appel, 2025) View paper
- Application-Specific Algorithms (1 papers)
  - [37] MULTI-LEVEL REGRESSION FOR NONLINEAR CON (BOUNDS, n.d.) View paper

Narrative

Core task: fine-grained gap-dependent regret bounds for model-free reinforcement learning. The field has evolved to address how quickly agents can learn near-optimal policies when the problem structure offers exploitable gaps between action values. The taxonomy reveals several major branches: tabular episodic methods that leverage optimism or posterior sampling to achieve instance-dependent guarantees; function approximation and structured settings that extend these ideas to large or continuous state spaces; specialized extensions covering risk-sensitivity, robustness, and non-stationarity; multi-agent and federated scenarios; and foundational frameworks that unify exploration strategies across diverse problem classes. Works such as Regret-Optimal Q-Learning[4] and Reference-Advantage Decomposition[7] exemplify how UCB-based optimistic algorithms refine gap-dependent analyses in tabular episodic MDPs, while Decision-Estimation Coefficients[6] and Instance-Optimality Theory[11] provide broader theoretical lenses. Meanwhile, branches on posterior sampling (e.g., Posterior Sampling Episodic[27]) and discounted infinite-horizon settings offer alternative algorithmic paradigms, and application-specific algorithms demonstrate how these principles transfer to real-world domains.

A particularly active line of research focuses on tightening regret bounds by exploiting finer notions of problem difficulty—moving beyond worst-case horizon or state-action counts to capture suboptimality gaps, variance, or cascaded structure (Cascaded Gaps[17], Sharp Variance Bounds[10]). Another contrasting direction explores robustness and adaptivity: handling adversarial perturbations (Robust Online Decisions[20]), non-stationary environments (Non-Stationary Episodic[29]), or federated learning constraints (Federated Q-Learning[9], Federated Learning Bounds[3]). Fine-Grained Q-Learning[0] sits squarely within the tabular episodic UCB-based cluster, emphasizing refined gap-dependent guarantees for model-free algorithms. It shares methodological kinship with Non-Asymptotic Gap-Dependent[28] and Tail Distribution Regret[25], which also pursue instance-specific bounds, but distinguishes itself by targeting finer-grained characterizations of the learning dynamics. This positioning highlights an ongoing effort to bridge worst-case and problem-dependent perspectives, offering practitioners tighter performance predictions when favorable problem structure is present.

Related Works in Same Category

The following **3 sibling papers** share the same taxonomy leaf node with the original paper:

1. Gap-dependent bounds for q-learning using reference-advantage decomposition

Authors: Zheng, Zhong, Zhang Haochen, Zhong Zheng, Xue, et al. (8 authors total) | **Year/Venue**: 2024 | **URL**: View paper

Abstract

We study the gap-dependent bounds of two important algorithms for on-policy Q-learning for finite-horizon episodic tabular Markov Decision Processes (MDPs): UCB-Advantage (Zhang et al. 2020) and Q-EarlySettled-Advantage (Li et al. 2021). UCB-Advantage and Q-

EarlySettled-Advantage improve upon the results based on Hoeffding-type bonuses and achieve the almost optimal $\sqrt{T}$-type regret bound in the worst-case scenario, where T is the total number of steps. However, the benign structures of ...

Relationship Analysis

Both papers belong to the UCB-Based Optimistic Algorithms category, focusing on achieving gap-dependent regret bounds for tabular episodic MDPs using upper confidence bound approaches. The candidate paper analyzes UCB-Advantage and Q-EarlySettled-Advantage algorithms that use variance estimators and reference-advantage decomposition, establishing gap-dependent bounds that are logarithmic in T. In contrast, the original paper develops a novel fine-grained analytical framework for UCB-Hoeffding and introduces ULCB-Hoeffding, achieving bounds with individual suboptimality gaps Δh(s,a) rather than relying solely on the global minimum gap Δmin, and also addresses non-UCB-based methods through a refined AMB algorithm.

---

2. Tail Distribution of Regret in Optimistic Reinforcement Learning

Authors: Sajad Khodadadian, Mehrdad Moharrami | **Year/Venue**: 2025 | **URL**: View paper

Abstract

We derive instance-dependent tail bounds for the regret of optimism-based reinforcement learning in finite-horizon tabular Markov decision processes with unknown transition dynamics. Focusing on a UCBVI-type algorithm, we characterize the tail distribution of the cumulative regret R_K over K episodes, rather than only its expectation or a single high-probability quantile. We analyze two natural exploration-bonus schedules: (i) a K-dependent scheme that explicitly incorporates the total num...

Relationship Analysis

Both papers belong to the UCB-Based Optimistic Algorithms category, using upper confidence bounds with Hoeffding bonuses to achieve gap-dependent regret in tabular episodic MDPs. The original paper focuses on establishing fine-grained gap-dependent regret bounds that depend on individual suboptimality gaps Δ_h(s,a) for both UCB-based and non-UCB-based algorithms, while the candidate paper characterizes the tail distribution of regret (P(R_K ≥ x)) for UCBVI-type algorithms, providing sub-Gaussian and sub-Weibull tail bounds rather than expected regret bounds. The key distinction is that the original paper derives fine-grained expected regret bounds with improved dependence on individual gaps, whereas the candidate paper analyzes the entire regret distribution beyond expectation.

---

3. Non-Asymptotic Gap-Dependent Regret Bounds for Tabular MDPs

Authors: Simchowitz, Max, Jamieson, Kevin, Max Simchowitz, et al. (6 authors total) | **Year/Venue**: 2019 | **URL**: View paper

Abstract

This paper establishes that optimistic algorithms attain gap-dependent and non-asymptotic logarithmic regret for episodic MDPs. In contrast to prior work, our bounds do not suffer a dependence on diameter-like quantities or ergodicity, and smoothly interpolate between the gap dependent logarithmic-regret, and the $\widetilde{\mathcal{O}}(\sqrt{HSAT})$-minimax rate. The key technique in our analysis is a novel "clipped" regret decomposition which applies to a broad family of recent optimistic alg...

Relationship Analysis

Both papers belong to the UCB-Based Optimistic Algorithms category, using upper confidence bounds to achieve gap-dependent regret in tabular episodic MDPs. The original paper (Q-Learning with Fine-Grained Gap-Dependent Regret) focuses on establishing fine-grained bounds with individual suboptimality gaps Δh(s,a) for model-free algorithms like UCB-Hoeffding, while the candidate paper (Non-Asymptotic Gap-Dependent Regret Bounds) establishes non-asymptotic logarithmic regret bounds for model-based optimistic algorithms like StrongEuler, with a coarser dependence on the minimum gap Δmin rather than individual gaps.

Contributions Analysis

Overall novelty summary. The paper establishes fine-grained gap-dependent regret bounds for model-free reinforcement learning in tabular episodic MDPs. It resides in the 'UCB-Based Optimistic Algorithms' leaf, which contains four papers total including the original work. This leaf sits within the broader 'Tabular Episodic MDP Algorithms with Gap-Dependent Bounds' branch, indicating a moderately populated research direction. The paper's focus on separating analysis of optimal versus suboptimal state-action pairs and introducing ULCB-Hoeffding represents an effort to refine existing UCB-based approaches in a well-established but still-active subfield.

The taxonomy reveals neighboring leaves addressing bootstrap methods and non-UCB optimistic approaches, both pursuing gap-dependent bounds through alternative mechanisms. The broader branch structure shows parallel efforts in function approximation, risk-sensitive objectives, and non-stationary environments. The paper's emphasis on fine-grained analysis connects it to theoretical foundations exploring instance-dependent complexity, while its UCB-based methodology distinguishes it from posterior sampling and bootstrap-based alternatives. The taxonomy's scope and exclude notes clarify that this work targets standard stationary episodic settings without function approximation or multi-agent constraints.

Among thirty candidates examined, the first contribution (novel analytical framework for UCB-based algorithms) showed no clear refutation across ten candidates, suggesting potential novelty in the separation technique for optimal and suboptimal pairs. The second contribution (ULCB-Hoeffding algorithm) similarly encountered no refuting candidates among ten examined, indicating the specific algorithmic design may be new. The third contribution (refined AMB algorithm) found one refutable candidate among ten examined, suggesting some overlap with prior AMB-related work. These statistics reflect a limited semantic search scope rather than exhaustive coverage of the field.

The analysis suggests moderate novelty within a focused research direction, with the UCB-based framework and ULCB-Hoeffding appearing more distinctive than the AMB refinement based on the limited search. The taxonomy positioning in a four-paper leaf indicates neither a highly crowded nor entirely sparse area. Readers should note that these assessments derive from top-thirty semantic matches and may not capture all relevant prior work, particularly in adjacent leaves or recent preprints outside the search scope.

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

Contribution 1: Novel fine-grained analytical framework for UCB-based algorithms

Description: The authors introduce a new analytical framework that separates the analysis of optimal and suboptimal state-action pairs, enabling the first fine-grained gap-dependent regret upper bound for UCB-Hoeffding and extending to ULCB-Hoeffding. This framework analyzes each state-action pair separately and establishes recursive relationships for cumulative weighted visitation counts.

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. Regret Analysis of Shrinking Horizon Model Predictive Control

URL: View paper

Brief Assessment

Shrinking Horizon MPC[42] focuses on model predictive control with computational constraints and regret analysis in control systems, not reinforcement learning. The candidate does not address UCB-based algorithms, state-action pair analysis, or gap-dependent regret bounds in tabular MDPs.

---

2. Near-optimal regret bounds for stochastic shortest path

URL: View paper

Brief Assessment

Stochastic Shortest Path[40] focuses on SSP problems with goal-state reaching objectives, not episodic tabular MDPs with fine-grained gap-dependent regret analysis separating optimal/suboptimal state-action pairs.

---

3. Asymptotically Optimal Regret for Black-Box Predict-then-Optimize

URL: View paper

Brief Assessment

Black-Box Predict-Optimize[41] addresses predict-then-optimize problems in supervised learning with binary decisions and reward prediction, not reinforcement learning regret analysis for UCB-based algorithms in MDPs.

---

4. The regret lower bound for communicating Markov Decision Processes

URL: View paper

Brief Assessment

Communicating MDPs[38] focuses on regret lower bounds for communicating MDPs in the average reward setting, not on fine-grained regret analysis for episodic tabular MDPs with UCB-based algorithms.

---

5. Warm-up free policy optimization: Improved regret in linear Markov decision processes

URL: View paper

Brief Assessment

Warm-Up Free[44] focuses on eliminating warm-up phases in policy optimization for linear MDPs, not on fine-grained gap-dependent regret analysis for tabular MDPs with UCB-based algorithms. The technical approaches are fundamentally different.

---

6. Near-optimal dynamic regret for adversarial linear mixture mdps

URL: View paper

Brief Assessment

Adversarial Linear Mixture[43] focuses on adversarial linear mixture MDPs with dynamic regret in non-stationary environments, not fine-grained gap-dependent regret analysis for tabular MDPs that separates optimal and suboptimal state-action pairs.

---

7. Fine-grained gap-dependent bounds for tabular mdps via adaptive multi-step bootstrap

URL: View paper

Brief Assessment

Adaptive Multi-Step Bootstrap[5] focuses on combining optimistic bootstrap with adaptive multi-step Monte Carlo rollout for non-UCB methods, not on developing a fine-grained analytical framework that separates optimal and suboptimal state-action pairs for UCB-based algorithms.

---

8. Test-time regret minimization in meta reinforcement learning

URL: View paper

Brief Assessment

Test-Time Regret Meta[39] focuses on meta reinforcement learning with task identification and test-time regret minimization, not on fine-grained gap-dependent regret analysis for single-task Q-learning algorithms.

---

9. Reinforcement learning can be more efficient with multiple rewards

URL: View paper

Brief Assessment

Multiple Rewards[45] focuses on multi-armed bandits and MDPs with multiple reward functions, not on fine-grained gap-dependent regret analysis for single-reward UCB algorithms. The paper's contribution is orthogonal to analyzing optimal vs. suboptimal state-action pairs separately.

---

10. Online Mixture of Experts: No-Regret Learning for Optimal Collective Decision-Making

URL: View paper

Brief Assessment

Mixture of Experts[46] focuses on expert-guided bandit learning for aggregating expert outputs in online settings, not on fine-grained regret analysis for reinforcement learning with state-action pair separation. The technical approaches are fundamentally different.

Contribution 2: ULCB-Hoeffding algorithm with fine-grained regret guarantees

Description: The authors propose ULCB-Hoeffding, a simplified UCB-based algorithm that removes the problematic multi-step bootstrapping from AMB while retaining the ULCB mechanism. This algorithm achieves fine-grained gap-dependent regret bounds and demonstrates improved empirical performance over AMB.

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. Federated UCBVI: Communication-Efficient Federated Regret Minimization with Heterogeneous Agents

URL: View paper

Brief Assessment

Federated UCBVI[50] focuses on federated multi-agent reinforcement learning with communication efficiency, not on fine-grained gap-dependent regret bounds for single-agent tabular MDPs. The technical approaches and problem settings are fundamentally different.

---

2. Efficient kernelized ucb for contextual bandits

URL: View paper

Brief Assessment

Kernelized UCB[53] focuses on computational efficiency for contextual bandits using kernel approximations, not on fine-grained gap-dependent regret bounds for tabular MDPs. The technical approaches and problem settings are fundamentally different.

---

3. Fast and regret optimal best arm identification: Fundamental limits and low-complexity algorithms

URL: View paper

Brief Assessment

Fast Best Arm[49] focuses on best arm identification in multi-armed bandits with dual objectives of quick commitment and reward maximization, not on fine-grained gap-dependent regret bounds for Q-learning in episodic tabular MDPs.

---

4. Breaking the sample complexity barrier to regret-optimal model-free reinforcement learning

URL: View paper

Brief Assessment

Breaking Sample Complexity[56] focuses on achieving regret-optimal model-free RL with memory efficiency (space complexity O(SAH)) and improved sample complexity requirements. The candidate does not address fine-grained gap-dependent regret bounds or the ULCB mechanism combined with simplified structure that removes multi-step bootstrapping while retaining optimism principles.

---

5. Non-stationary Reinforcement Learning under General Function Approximation

URL: View paper

Brief Assessment

Non-Stationary Function Approximation[52] focuses on non-stationary MDPs with general function approximation and sliding window mechanisms, not on fine-grained gap-dependent regret bounds for tabular MDPs with UCB-based algorithms.

---

6. A domain-shrinking based Bayesian optimization algorithm with order-optimal regret performance

URL: View paper

Brief Assessment

Domain-Shrinking Bayesian[54] focuses on Bayesian optimization in continuous domains using Gaussian processes, not Q-learning in episodic tabular MDPs. The technical approaches are fundamentally different.

---

7. Comparative analysis of Sliding Window UCB and Discount Factor UCB in non-stationary environments: A Multi-Armed Bandit approach

URL: View paper

Brief Assessment

Sliding Window UCB[48] focuses on non-stationary multi-armed bandit problems with sliding window and discount factor approaches, not on episodic tabular MDPs with fine-grained gap-dependent regret bounds for reinforcement learning.

---

8. Multiagent Online Source Seeking Using Bandit Algorithm

URL: View paper

Brief Assessment

Multiagent Source Seeking[51] addresses a fundamentally different problem domain (online source-seeking with multiagent systems in dynamical environments) rather than episodic tabular MDPs. The D-UCB mechanism is designed for distributed task planning in source-seeking, not for fine-grained regret analysis in reinforcement learning.

---

9. Data-Driven Upper Confidence Bounds with Near-Optimal Regret for Heavy-Tailed Bandits

URL: View paper

Brief Assessment

Heavy-Tailed Bandits[55] focuses on multi-armed bandits with heavy-tailed reward distributions and proposes a data-driven UCB method (RMM-UCB) that does not require moment parameters. The original paper addresses Q-learning in episodic tabular MDPs with a simplified UCB-based algorithm (ULCB-Hoeffding) that removes multi-step bootstrapping. These are fundamentally different problem settings and algorithmic approaches.

---

10. Regret Guarantees for a UCB-based Algorithm for Volatile Combinatorial Bandits

URL: View paper

Brief Assessment

Volatile Combinatorial Bandits[47] addresses a different problem setting (combinatorial bandits with volatile/unavailable arms) and proposes CV-UCB algorithm. The technical focus on availability constraints and combinatorial action spaces differs fundamentally from ULCB-Hoeffding's episodic tabular MDP setting with ULCB mechanisms.

Contribution 3: Refined AMB algorithm with rigorous fine-grained analysis

Description: The authors identify algorithmic and analytical issues in the original AMB algorithm and propose Refined AMB, which removes improper truncations, rigorously proves unbiasedness of multi-step bootstrapping estimators, ensures martingale difference conditions hold, and establishes the first valid fine-grained gap-dependent regret bound for a non-UCB-based method.

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. Instance-dependent regret bounds for learning two-player zero-sum games with bandit feedback

URL: View paper

Brief Assessment

Two-Player Zero-Sum[57] focuses on instance-dependent regret bounds for two-player zero-sum games with bandit feedback, not on reinforcement learning in episodic tabular MDPs. The technical approaches differ fundamentally: the candidate addresses game-theoretic equilibria while the original addresses Q-learning in MDPs.

---

2. Gap-dependent bounds for two-player markov games

URL: View paper

Brief Assessment

Two-Player Markov Games[62] focuses on two-player turn-based stochastic games with Nash equilibrium learning, not single-agent RL with fine-grained gap-dependent bounds for non-UCB methods like AMB.

---

3. Instance-Dependent Complexity of Contextual Bandits and Reinforcement Learning: A Disagreement-Based Perspective

URL: View paper

Brief Assessment

Disagreement-Based Complexity[32] focuses on contextual bandits and reinforcement learning with disagreement-based complexity measures, not on fine-grained gap-dependent regret bounds for Q-learning algorithms like AMB. The candidate addresses different technical problems in a different algorithmic framework.

---

4. Reinforcement Learning from Human Feedback with Active Queries

URL: View paper

Brief Assessment

Active Queries RLHF[61] addresses alignment of LLMs with human preferences using contextual dueling bandits and active learning, which is fundamentally different from the original paper's focus on fine-grained gap-dependent regret bounds for model-free Q-learning in episodic tabular MDPs. The candidate does not address multi-step bootstrapping, martingale difference conditions, or Q-function estimation issues in reinforcement learning.

---

5. Optimistic pac reinforcement learning: the instance-dependent view

URL: View paper

Brief Assessment

Optimistic PAC[58] focuses on PAC reinforcement learning with optimistic sampling rules and does not address the specific algorithmic issues in AMB (improper truncations, multi-step bootstrapping unbiasedness, martingale difference conditions) that the original paper identifies and corrects.

---

6. Fine-grained gap-dependent bounds for tabular mdps via adaptive multi-step bootstrap

URL: View paper

Prior Art Analysis

Adaptive Multi-Step Bootstrap[5] presents the original AMB algorithm that the ORIGINAL paper claims to refine. The candidate paper describes the same core algorithmic components (multi-step bootstrapping, ULCB mechanism, action elimination) and provides theoretical analysis establishing fine-grained gap-dependent bounds. The ORIGINAL paper's claim to have 'identified algorithmic and analytical issues' and proposed a 'refined version' is directly challenged by the existence of this prior work presenting the AMB algorithm with its multi-step bootstrapping approach and gap-dependent analysis.

Evidence

Evidence 1 - **Rationale**: The candidate paper introduces the original AMB algorithm with multi-step bootstrapping, which the ORIGINAL paper claims to be refining. This demonstrates that AMB with multi-step bootstrapping existed prior to the ORIGINAL paper's claimed refinement. - **Original**: we revisit the only known algorithm amb, and identify two key issues in its algorithm design and analysis: improper truncation in the q-updates and violation of the martingale difference condition in its concentration argument. we propose a refined version of amb that addresses these issues - **Candidate**: this paper presents a new model-free algorithm for episodic finite-horizon markov decision processes (mdp), adaptive multi-step bootstrap (amb), which enjoys a stronger gap-dependent regret bound. the first innovation is to estimate the optimalq-function by combining an optimistic bootstrap with an ad...

Evidence 2 - **Rationale**: The candidate establishes that AMB already achieved gap-dependent bounds, contradicting the ORIGINAL paper's claim to be the first to establish such bounds for a non-UCB method through their refinement. - **Original**: in the non-ucb-based setting, we revisit the only known algorithm amb, and identify two key issues in its algorithm design and analysis: improper truncation in the q-updates and violation of the martingale difference condition in its concentration argument. - **Candidate**: we show when each state has a unique optimal action, amb achieves a gapdependent regret bound that only scales with the sum of the inverse of the sub-optimality gaps. in contrast, simchowitz and jamieson (2019) showed all upper-confidence-bound (ucb) algorithms suffer an additional ω (s Δmin) regre...

Evidence 3 - **Rationale**: The candidate describes the core algorithmic structure including upper/lower bounds and action elimination that the ORIGINAL paper claims to be refining, showing these mechanisms existed in the original AMB. - **Original**: removes improper truncations in theq-updates, (ii) rigorously proves that the estimators induced by multi-step bootstrapping form an unbiased estimate of the optimal q-function, (iii) ensures the martingale difference condition holds - **Candidate**: our algorithm maintains valid upper bounds and lower bounds of the q-function at every episode k, denoted by qk(x,a) and qk(x,a), respectively. given these bounds, for every state x, it maintains a set of candidate optimal actions, denoted by ak(x), by eliminating every action a whose q-value upper ...

---

7. Regret-Optimal Q-Learning with Low Cost for Single-Agent and Federated Reinforcement Learning

URL: View paper

Brief Assessment

Regret-Optimal Q-Learning[4] focuses on achieving near-optimal regret with low burn-in costs and logarithmic switching/communication costs in single-agent and federated RL settings. It does not address fine-grained gap-dependent regret bounds for non-UCB-based methods or the specific algorithmic issues in AMB that the original paper identifies and resolves.

---

8. Waypoint-Based Reinforcement Learning for Robot Manipulation Tasks

URL: View paper

Brief Assessment

Waypoint-Based Manipulation[60] focuses on robot manipulation tasks using waypoints and multi-armed bandits, not on fine-grained gap-dependent regret bounds for Q-learning or reinforcement learning algorithms in tabular MDPs. The technical domains are fundamentally different.

---

9. A tighter problem-dependent regret bound for risk-sensitive reinforcement learning

URL: View paper

Brief Assessment

Risk-Sensitive Regret[1] focuses on risk-sensitive RL with exponential utility in episodic MDPs, not on fine-grained gap-dependent regret bounds for general model-free RL methods like the original paper's Refined AMB.

---

10. More Benefits of Being Distributional: Second-Order Bounds for Reinforcement Learning

URL: View paper

Brief Assessment

Distributional Second-Order[59] focuses on distributional RL with second-order bounds and variance-based analysis, not on fine-grained gap-dependent regret bounds for non-UCB-based methods like AMB. The technical approaches are fundamentally different.

Appendix: Text Similarity Detection

Textual similarity detection checked 32 papers and found 3 similarity segment(s) across 1 paper(s).

The following **1 paper(s)** were detected to have high textual similarity with the original paper. These may represent different versions of the same work, duplicate submissions, or papers with substantial textual overlap. Readers are advised to verify these relationships independently.

1. Regret-Optimal Q-Learning with Low Cost for Single-Agent and Federated Reinforcement Learning

Detected in: Contribution: contribution_3

⚠ **Note**: This paper shows substantial textual similarity with the original paper. It may be a different version, a duplicate submission, or contain significant overlapping content. Please review carefully to determine the nature of the relationship.

References

- [0] Q-Learning with Fine-Grained Gap-Dependent Regret View paper
- [1] A tighter problem-dependent regret bound for risk-sensitive reinforcement learning View paper
- [2] Information-directed pessimism for offline reinforcement learning View paper
- [3] Gap-Dependent Bounds for Federated -learning View paper
- [4] Regret-Optimal Q-Learning with Low Cost for Single-Agent and Federated Reinforcement Learning View paper
- [5] Fine-grained gap-dependent bounds for tabular mdps via adaptive multi-step bootstrap View paper
- [6] Unified algorithms for RL with Decision-Estimation Coefficients: PAC, reward-free, preference-based learning and beyond View paper
- [7] Gap-dependent bounds for q-learning using reference-advantage decomposition View paper
- [8] Efficient Restarts in Non-Stationary Model-Free Reinforcement Learning View paper
- [9] Gap-Dependent Regret for Federated Q-Learning View paper
- [10] Sharp Variance-Dependent Bounds in Reinforcement Learning: Best of Both Worlds in Stochastic and Deterministic Environments View paper
- [11] Instance-optimality in interactive decision making: Toward a non-asymptotic theory View paper
- [12] VOL: Towards Optimal Regret in Model-free RL with Nonlinear Function Approximation View paper
- [13] Politex: Regret bounds for policy iteration using expert prediction View paper
- [14] Randomized exploration for reinforcement learning with multinomial logistic function approximation View paper
- [15] Hybrid reinforcement learning breaks sample size barriers in linear mdps View paper
- [16] Adaptive discretization in online reinforcement learning View paper
- [17] Cascaded Gaps: Towards Gap-Dependent Regret for Risk-Sensitive Reinforcement Learning View paper
- [18] Logarithmic regret for reinforcement learning with linear function approximation View paper
- [19] Beyond Value-Function Gaps: Improved Instance-Dependent Regret Bounds for Episodic Reinforcement Learning View paper
- [20] Regret Bounds for Robust Online Decision Making View paper
- [21] The best of both worlds: Reinforcement learning with logarithmic regret and policy switches View paper
- [22] Model-Free Nonstationary Reinforcement Learning: Near-Optimal Regret and Applications in Multiagent Reinforcement Learning and Inventory Control View paper
- [23] Can Q-learning be improved with advice? View paper
- [24] Model-free online learning in unknown sequential decision making problems and games View paper
- [25] Tail Distribution of Regret in Optimistic Reinforcement Learning View paper
- [26] Advances in Inverse Problems and Decision-making under Uncertainty View paper
- [27] A Provably Efficient Model-Free Posterior Sampling Method for Episodic Reinforcement Learning View paper
- [28] Non-Asymptotic Gap-Dependent Regret Bounds for Tabular MDPs View paper
- [29] Near-Optimal Regret Bounds for Model-Free RL in Non-Stationary Episodic MDPs View paper
- [30] Regret-Optimal Model-Free Reinforcement Learning for Discounted MDPs with Short Burn-In Time View paper
- [31] Generalized Bandit Regret Minimizer Framework in Imperfect Information Extensive-Form Game View paper
- [32] Instance-Dependent Complexity of Contextual Bandits and Reinforcement Learning: A Disagreement-Based Perspective View paper
- [33] Model-Free Non-Stationary RL: Near-Optimal Regret and Applications in Multi-Agent RL and Inventory Control View paper
- [34] Regret Bound Balancing and Elimination for Model Selection in Bandits and RL View paper
- [35] Online Regret Bounds for Undiscounted Continuous Reinforcement Learning View paper
- [36] Efficient Reinforcement Learning via Initial Pure Exploration View paper
- [37] MULTI-LEVEL REGRESSION FOR NONLINEAR CON View paper
- [38] The regret lower bound for communicating Markov Decision Processes View paper
- [39] Test-time regret minimization in meta reinforcement learning View paper
- [40] Near-optimal regret bounds for stochastic shortest path View paper
- [41] Asymptotically Optimal Regret for Black-Box Predict-then-Optimize View paper
- [42] Regret Analysis of Shrinking Horizon Model Predictive Control View paper
- [43] Near-optimal dynamic regret for adversarial linear mixture mdps View paper
- [44] Warm-up free policy optimization: Improved regret in linear Markov decision processes View paper
- [45] Reinforcement learning can be more efficient with multiple rewards View paper
- [46] Online Mixture of Experts: No-Regret Learning for Optimal Collective Decision-Making View paper
- [47] Regret Guarantees for a UCB-based Algorithm for Volatile Combinatorial Bandits View paper
- [48] Comparative analysis of Sliding Window UCB and Discount Factor UCB in non-stationary environments: A Multi-Armed Bandit approach View paper
- [49] Fast and regret optimal best arm identification: Fundamental limits and low-complexity algorithms View paper
- [50] Federated UCBVI: Communication-Efficient Federated Regret Minimization with Heterogeneous Agents View paper
- [51] Multiagent Online Source Seeking Using Bandit Algorithm View paper
- [52] Non-stationary Reinforcement Learning under General Function Approximation View paper
- [53] Efficient kernelized ucb for contextual bandits View paper
- [54] A domain-shrinking based Bayesian optimization algorithm with order-optimal regret performance View paper
- [55] Data-Driven Upper Confidence Bounds with Near-Optimal Regret for Heavy-Tailed Bandits View paper
- [56] Breaking the sample complexity barrier to regret-optimal model-free reinforcement learning View paper
- [57] Instance-dependent regret bounds for learning two-player zero-sum games with bandit feedback View paper

- [58] Optimistic pac reinforcement learning: the instance-dependent view View paper
- [59] More Benefits of Being Distributional: Second-Order Bounds for Reinforcement Learning View paper
- [60] Waypoint-Based Reinforcement Learning for Robot Manipulation Tasks View paper
- [61] Reinforcement Learning from Human Feedback with Active Queries View paper
- [62] Gap-dependent bounds for two-player markov games View paper