Novelty Assessment Report

Paper: XQC: Well-conditioned Optimization Accelerates Deep Reinforcement Learning

PDF URL: https://openreview.net/pdf?id=tx1ZvypKqS

Venue: ICLR 2026 Conference Submission

Year: 2026

Report Generated: 2026-01-01

Abstract

Sample efficiency is a central property of effective deep reinforcement learning algorithms. Recent work has improved this through added complexity, such as larger models, exotic network architectures, and more complex algorithms, which are typically motivated purely by empirical performance. We take a more principled approach by focusing on the optimization landscape of the critic network. Using the eigenspectrum and condition number of the critic's Hessian, we systematically investigate the impact of common architectural design decisions on training dynamics. Our analysis reveals that a novel combination of batch normalization (BN), weight normalization (WN), and a distributional cross-entropy (CE) loss produces condition numbers orders of magnitude smaller than baselines. This combination also naturally bounds gradient norms, a property critical for maintaining a stable effective learning rate under non-stationary targets and bootstrapping. Based on these insights, we introduce XQC: a well-motivated, sample-efficient deep actor-critic algorithm built upon soft actor-critic that embodies these optimization-aware principles. We achieve state-of-the-art sample efficiency across 55 proprioception and 15 vision-based continuous control tasks, all while using significantly fewer parameters than competing methods.

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: **Improving Sample Efficiency in Deep Reinforcement Learning through Well-Conditioned Critic Optimization**

A total of **50 papers** were analyzed and organized into a taxonomy with **36 categories**.

Taxonomy Overview

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

- **Critic Optimization and Value Function Learning**
- **Representation Learning for Sample Efficiency**
- **Policy Learning and Actor-Critic Integration**
- **Exploration and Data Collection Strategies**
- **Experience Replay and Memory Management**
- **Transfer Learning and Knowledge Reuse**
- **Scaling and Computational Efficiency**
- **Domain-Specific Applications and Adaptations**
- **Specialized Techniques and Constraints**
- **Surveys, Frameworks, and Theoretical Foundations**

Complete Taxonomy Tree

- Improving Sample Efficiency in Deep Reinforcement Learning through Well-Conditioned Critic Optimization Survey Taxonomy
- Critic Optimization and Value Function Learning
  - Critic Architecture and Conditioning ★ (3 papers)
  - [0] XQC: Well-conditioned Optimization Accelerates Deep Reinforcement Learning (Anon et al., 2026) View paper
  - [3] Boosting On-Policy Actorâ□□Critic With Shallow Updates in Critic (Luntong Li, 2024) View paper
  - [12] CTD4 - A Deep Continuous Distributional Actor-Critic Agent with a Kalman Fusion of Multiple Critics (David Valencia, 2024) View paper
  - Critic Update Mechanisms and Temporal Difference Learning (3 papers)
  - [2] Efficient offline reinforcement learning: The critic is critical (Jelley, 2024) View paper
  - [15] Iterated Q-Network: Beyond One-Step Bellman Updates in Deep Reinforcement Learning (Vincent Theo, 2024) View paper
  - [48] Data Efficient Deep Reinforcement Learning With Action-Ranked Temporal Difference Learning (Qi Liu, 2024) View paper
  - Multi-Step and Iterated Value Learning (1 papers)
  - [36] A novel multi-step Q-learning method to improve data efficiency for deep reinforcement learning (Yinlong Yuan, 2019) View paper
  - Value Function Generalization and Robustness (1 papers)
  - [11] Rethinking value function learning for generalization in reinforcement learning (Moon, 2022) View paper
- Representation Learning for Sample Efficiency
  - Contrastive and Self-Supervised Representation Learning (2 papers)
  - [1] On the data-efficiency with contrastive image transformation in reinforcement learning (S Liu, 2023) View paper
  - [7] Improving sample-efficiency of model-free reinforcement learning algorithms on image inputs with representation learning (M Guberina, 2022) View paper
  - Value-Consistent and Planning-Oriented Representations (1 papers)
  - [23] Value-consistent representation learning for data-efficient reinforcement learning (Yue Yang, 2023) View paper

  - Geometric and Structured Embeddings (1 papers)
  - [6] Simplicial embeddings improve sample efficiency in actor-critic agents (Obando-Ceron, 2025) View paper
  - Theoretical Foundations of Representation Sufficiency (1 papers)
  - [14] Is a good representation sufficient for sample efficient reinforcement learning? (Du, 2019) View paper
- Policy Learning and Actor-Critic Integration
  - On-Policy and Off-Policy Integration (1 papers)
  - [24] Interpolated Policy Gradient: Merging On-Policy and Off-Policy Gradient Estimation for Deep Reinforcement Learning (Gu Shixiang, 2017) View paper
  - Actor-Critic Coordination and Feedback Mechanisms (1 papers)
  - [49] A Feedback-based Decision-Making Mechanism for Actor-Critic Deep Reinforcement Learning (Guang Yang, 2024) View paper
  - Policy Gradient and Advantage Estimation Methods (1 papers)
  - [47] AM-PPO:(Advantage) Alpha-Modulation with Proximal Policy Optimization (Sane, 2025) View paper
- Exploration and Data Collection Strategies
  - Curiosity-Driven and Intrinsic Motivation (1 papers)
  - [33] Investigating Sample Efficient Deep Reinforcement Learning (Mavor-Parker, 2025) View paper
  - Satisficing and Adaptive Exploration (2 papers)
  - [44] Adaptive Exploration for Data-Efficient General Value Function Evaluations (Josiah Hanna, 2024) View paper
  - [45] Satisficing Exploration for Deep Reinforcement Learning (Arumugam, 2024) View paper
  - Hierarchical and Goal-Conditioned Exploration (1 papers)
  - [20] Guided cooperation in hierarchical reinforcement learning via model-based rollout (Haoran Wang, 2024) View paper
- Experience Replay and Memory Management
  - Prioritized and Enhanced Replay Mechanisms (1 papers)
  - [13] Efficient deep reinforcement learning-enabled recommendation (Guangyao Pang, 2022) View paper
  - Hindsight Experience Replay and Goal Relabeling (1 papers)
  - [17] Learning and Reusing Primitive Behaviours to Improve Hindsight Experience Replay Sample Efficiency (Francisco Roldan Sanchez, 2023) View paper
  - Memory-Efficient Replay and Knowledge Consolidation (1 papers)
  - [28] Memory-efficient Reinforcement Learning with Value-based Knowledge Consolidation (Lan, 2022) View paper
- Transfer Learning and Knowledge Reuse
  - Value Function Initialization and Transfer (1 papers)
  - [40] Value Function Initialization for Knowledge Transfer and Jump-start in Deep Reinforcement Learning (Mehimeh, 2025) View paper
  - Video and Observation-Only Pre-training (3 papers)
  - [27] Robotic Offline RL from Internet Videos via Value-Function Learning (Chethan Bhateja, 2024) View paper
  - [38] Robotic Offline RL from Internet Videos via Value-Function Pre-Training (Bhateja, 2023) View paper
  - [50] ViVa: Video-Trained Value Functions for Guiding Online RL from Diverse Data (Dashora, 2025) View paper
  - Policy Transfer and Lifelong Learning (1 papers)
  - [8] Policy Correction and State-Conditioned Action Evaluation for Few-Shot Lifelong Deep Reinforcement Learning (Meng Xu, 2024) View paper
  - Meta-Learning and Learning-to-Learn (1 papers)
  - [19] Learning to learn: Meta-critic networks for sample efficient learning (Sung, 2017) View paper
- Scaling and Computational Efficiency
  - Scaling Laws and Predictability (1 papers)
  - [37] Value-Based Deep RL Scales Predictably (Rybkin, 2025) View paper
  - Compute-Optimal Training and Resource Allocation (1 papers)
  - [42] Compute-Optimal Scaling for Value-Based Deep RL (Rybkin, 2025) View paper
  - Length-Aware and Efficiency-Focused Optimization (1 papers)
  - [41] Efficient RL Training for Reasoning Models via Length-Aware Optimization (Xie Tian, 2025) View paper
- Domain-Specific Applications and Adaptations
  - Robotic Manipulation and Control (1 papers)
  - [4] Data-efficient deep reinforcement learning for dexterous manipulation (Popov, 2017) View paper
  - Robotic Navigation and Path Planning (1 papers)
  - [34] Robot Path Planning in Crowd Environments via Dynamic Interaction Modeling and Deep Reinforcement Learning (Yixiang Zhang, 2025) View paper
  - Multi-Agent and Cooperative Systems (1 papers)
  - [21] Investigating effects of centralized learning decentralized execution on team coordination in the level based foraging environment as a sequential social dilemma (Peter Atrazhev, 2022) View paper
  - Recommendation Systems (1 papers)
  - [26] Recommendation of deep reinforcement learning based on value function considering error reduction (JinLian Zhou, 2025) View paper
  - Control Systems and Optimization (2 papers)
  - [9] Adaptive critic learning for approximate optimal event-triggered tracking control of nonlinear systems with prescribed performances (Tengda Wang, 2023) View paper
  - [43] A nonlinear robust control method based on distributed deep reinforcement learning (Tianlei Wang, 2025) View paper
  - Edge Computing and Resource-Constrained Deployment (1 papers)
  - [22] Deep Reinforcement Learning Based on Parked Vehicles-Assisted for Task Offloading in Vehicle Edge Computing (Bingxin Wang, 2023) View paper
  - Financial Trading and Decision Systems (1 papers)
  - [16] Deep Reinforcement Learning Algorithms for Profitable Stock Trading Strategies (Abhiman Sakilam, 2024) View paper
  - Architectural and Creative Design (1 papers)
  - [46] Exploration of DeepSeek System Supported by Deep Reinforcement Learning in Architectural Creative Design (Zhang, 2025) View paper
  - Multi-Objective Optimization Applications (1 papers)

    - [25] Investigating the multi-objective optimization of quality and efficiency using deep reinforcement learning (Zhenhui Wang, 2022) View paper
- Specialized Techniques and Constraints
    - Safe Reinforcement Learning (1 papers)
    - [18] Constraint-conditioned policy optimization for versatile safe reinforcement learning (Yao, 2023) View paper
    - Data Valuation and Selection (1 papers)
    - [5] Data valuation using reinforcement learning (Jinsung Yoon, 2020) View paper
- Surveys, Frameworks, and Theoretical Foundations (7 papers)
    - [10] Data-efficient controller tuning and reinforcement learning (FrÃ¶hlich, 2022) View paper
    - [29] Advancing data-efficiency in reinforcement learning (Madjiheurem, 2022) View paper
    - [30] Towards Data-efficient AI: Theoretical analysis and experimental validation of new exploration algorithms for Reinforcement Learning (Kayal, 2025) View paper
    - [31] Sample-efficient deep reinforcement learning for control, exploration and safety (Flet-Berliac, 2021) View paper
    - [32] Efficient Data Usage for Planning and Reinforcement Learning (Xu, 2025) View paper
    - [35] Bridging Exploration and General Function Approximation in Reinforcement Learning: Provably Efficient Kernel and Neural Value Iterations (Zhuoran Yang, 2020) View paper
    - [39] Interaction-Efficient Reinforcement Learning: Matching the Real World Data Availability (Jiang, 2025) View paper

Narrative

Core task: improving sample efficiency in deep reinforcement learning through well-conditioned critic optimization. The field addresses the challenge of learning effective policies from limited environment interactions by refining how value functions are estimated and how representations are learned. The taxonomy reveals a rich structure spanning ten major branches. Critic Optimization and Value Function Learning focuses on architectural choices and conditioning strategies that stabilize temporal-difference updates, while Representation Learning for Sample Efficiency explores how to extract compact, task-relevant features from high-dimensional observations. Policy Learning and Actor-Critic Integration examines the interplay between policy updates and value estimation, and Exploration and Data Collection Strategies investigates how agents can gather informative experiences. Experience Replay and Memory Management considers how stored transitions can be reused effectively, and Transfer Learning and Knowledge Reuse looks at leveraging prior knowledge across tasks. Scaling and Computational Efficiency addresses resource constraints, Domain-Specific Applications and Adaptations tailors methods to particular problem settings, Specialized Techniques and Constraints handles safety and multi-objective scenarios, and Surveys, Frameworks, and Theoretical Foundations provides overarching perspectives.

Several active lines of work highlight contrasting emphases and open questions. One thread investigates critic architecture and conditioning—how the structure and update rules of value networks influence learning stability and sample complexity. For instance, Shallow Critic Updates[3] and CTD4 Kalman Fusion[12] explore different mechanisms for controlling critic complexity and noise. Another thread examines representation quality, asking whether good features alone suffice for efficiency or whether joint optimization of critics and encoders is essential, as debated in works like Good Representation Sufficient[14] and Value-Consistent Representation[23]. XQC[0] sits squarely within the critic architecture and conditioning cluster, emphasizing well-conditioned optimization to reduce variance in value estimates. Compared to Shallow Critic Updates[3], which limits network depth to improve conditioning, XQC[0] appears to pursue complementary regularization strategies that directly shape the critic's Hessian or gradient landscape, aiming for smoother, more stable learning dynamics without sacrificing representational capacity.

Related Works in Same Category

The following **2 sibling papers** share the same taxonomy leaf node with the original paper:

1. Boosting On-Policy Actorâ□□Critic With Shallow Updates in Critic

Authors: Luntong Li, Yuanheng Zhu | **Year/Venue**: 2024 • IEEE Transactions on Neural Networks and Learning Systems | **URL**: View paper

Abstract

Deep reinforcement learning (DRL) benefits from the representation power of deep neural networks (NNs), to approximate the value function and policy in the learning process. Batch reinforcement learning (BRL) benefits from stable training and data efficiency with fixed representation and enjoys solid theoretical analysis. This work proposes least-squares deep policy gradient (LSDPG), a hybrid approach that combines least-squares reinforcement learning (RL) with online DRL to achieve the best of ...

Relationship Analysis

Both papers belong to the Critic Architecture and Conditioning category, focusing on improving critic training stability through architectural innovations. While XQC addresses critic conditioning through batch normalization, weight normalization, and distributional cross-entropy loss to improve the optimization landscape, LSDPG takes a fundamentally different approach by combining least-squares temporal difference methods with deep learning, treating the critic as a linear combination of learned representations and performing regularized LSTD updates. The key distinction is that XQC focuses on continuous optimization dynamics and Hessian conditioning, whereas LSDPG employs batch-style least-squares methods with periodic representation updates in an on-policy setting.

2. CTD4 - A Deep Continuous Distributional Actor-Critic Agent with a Kalman Fusion of Multiple Critics

Authors: David Valencia, Henry P. Williams, Yuning Xing, Trevor Gee, Bruce A. MacDonald, et al. (6 authors total) | **Year/Venue**: 2024 • AAAI Conference on Artificial Intelligence | **URL**: View paper

Abstract

Categorical Distributional Reinforcement Learning (CDRL) has demonstrated superior sample efficiency in learning complex tasks compared to conventional Reinforcement Learning (RL) approaches. However, the practical application of CDRL is encumbered by challenging projection steps, detailed parameter tuning, and domain knowledge. This paper addresses these challenges by introducing a pioneering Continuous Distributional Model-Free RL algorithm tailored for continuous action spaces. The proposed a...

Relationship Analysis

Both papers belong to the Critic Architecture and Conditioning category, focusing on architectural innovations to improve critic training stability. They share an overlapping focus on distributional methods for value function learning, with both employing categorical distributional critics and cross-entropy losses to improve optimization conditioning. However, the original paper (XQC) emphasizes the synergistic combination of batch normalization, weight normalization, and distributional critics with extensive Hessian analysis to achieve well-conditioned optimization, while the candidate paper (CTD4) introduces a continuous distributional model with Kalman fusion of multiple critics to address overestimation bias, representing different architectural solutions within the same problem space.

Contributions Analysis

Overall novelty summary. The paper introduces XQC, a sample-efficient actor-critic algorithm that combines batch normalization, weight normalization, and distributional cross-entropy loss to improve critic conditioning. It resides in the 'Critic Architecture and Conditioning' leaf, which contains only three papers total, indicating a relatively sparse research direction within the broader taxonomy. This leaf focuses specifically on architectural innovations for critic stability, distinguishing it from adjacent leaves that address update mechanisms or multi-step learning. The small population suggests this particular angle—optimizing conditioning through architectural choices—remains underexplored compared to other critic optimization strategies.

The taxonomy reveals that critic optimization branches into four distinct leaves: architecture/conditioning, update mechanisms, multi-step learning, and generalization/robustness. XQC's focus on Hessian conditioning and normalization techniques positions it closest to architectural concerns, while neighboring leaves like 'Critic Update Mechanisms' (containing TD learning variants) and 'Multi-Step Value Learning' pursue complementary angles. The broader 'Representation Learning' branch (four leaves, multiple papers per leaf) represents a parallel research thrust emphasizing feature quality over critic conditioning. XQC's approach diverges by treating conditioning as a first-order concern rather than relying primarily on representation improvements or update rule modifications.

Among 21 candidates examined across three contributions, the XQC algorithm itself shows overlap with prior work: 10 candidates examined, 2 refutable. The Hessian eigenvalue analysis contribution appears more novel (1 candidate, 0 refutable), while the cross-entropy versus squared error analysis examined 10 candidates with none refutable. This suggests the algorithmic combination may have precedent in the limited search scope, but the specific conditioning analysis and loss comparison appear less directly addressed. The modest search scale (21 total candidates, not hundreds) means these findings reflect top semantic matches rather than exhaustive coverage, leaving room for additional related work outside this scope.

Given the sparse taxonomy leaf and limited search scope, the work appears to occupy a relatively underexplored niche within critic optimization. The conditioning-focused perspective distinguishes it from update-mechanism or representation-centric approaches, though the algorithmic contribution shows some overlap among examined candidates. The analysis contributions (Hessian eigenvalues, loss comparison) seem less directly refuted within this search scope, suggesting potential novelty in the diagnostic framework even if the final algorithm builds on established components.

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

Contribution 1: XQC algorithm for sample-efficient deep reinforcement learning

Description: The authors propose XQC, a deep actor-critic algorithm that extends soft actor-critic by combining batch normalization, weight normalization, and a distributional cross-entropy loss to create a well-conditioned optimization landscape. This design achieves state-of-the-art sample efficiency across 70 continuous control tasks while using significantly fewer parameters than competing methods.

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. Striving for simplicity and performance in off-policy DRL: Output normalization and non-uniform sampling

URL: View paper

Brief Assessment

Output Normalization Sampling[64] focuses on output normalization and non-uniform sampling (ERE) for off-policy RL, not on creating well-conditioned optimization landscapes through batch normalization, weight normalization, and distributional cross-entropy loss as XQC does. The candidate addresses bounded action spaces and replay buffer sampling, which are orthogonal concerns to XQC's optimization-centric approach.

2. Combining policy gradient and Q-learning

URL: View paper

Brief Assessment

Combining PG Q-Learning[62] focuses on combining policy gradient with Q-learning methods for data efficiency, while the original paper proposes XQC with batch normalization, weight normalization, and distributional cross-entropy loss for well-conditioned optimization. The candidate's limited context does not demonstrate prior work on the specific architectural combination or optimization landscape analysis that forms the core of XQC's novelty claim.

3. Optimistic Actor-Critic with Parametric Policies: Unifying Sample Efficiency and Practicality

URL: View paper

Brief Assessment

Optimistic Parametric Policies[70] focuses on actor-critic methods with parametric policies for episodic linear MDPs using Langevin Monte Carlo for exploration, whereas XQC addresses sample efficiency through optimization landscape conditioning via batch normalization, weight normalization, and distributional cross-entropy loss in continuous control tasks.

4. Personalized Recommendation System Based on Deep Reinforcement Learning

URL: View paper

Brief Assessment

Personalized Recommendation[69] focuses on recommendation systems using actor-critic for user preference modeling, not on general continuous control tasks or optimization landscape conditioning through normalization techniques.

5. Autonomous Navigation of Mobile Robots in Complex Environments with Global Path Smoothing and Adaptive Local Control

URL: View paper

Brief Assessment

Global Smoothing Local[66] focuses on autonomous mobile robot navigation using RRT path planning combined with SAC for local control, not on developing novel deep actor-critic algorithms with normalization techniques for sample efficiency.

6. Optimizing Game Strategies with Deep Reinforcement Learning: A Framework for Intelligent Decision-Making

URL: View paper

Brief Assessment

Game Strategies Framework[71] focuses on sports game strategy optimization using deep soft actor-critic with moss growth optimization for NBA data, not on general sample-efficient deep RL with normalization techniques and distributional losses for continuous control tasks.

7. Crossq: Batch normalization in deep reinforcement learning for greater sample efficiency and simplicity

URL: View paper

Prior Art Analysis

CrossQ Batch Normalization[65] demonstrates that a similar approach combining batch normalization with actor-critic methods for sample-efficient deep RL was proposed earlier. Both papers propose algorithms that extend soft actor-critic by incorporating batch normalization to achieve state-of-the-art sample efficiency in continuous control tasks. CrossQ[65] explicitly states it 'matches or surpasses current state-of-the-art methods in terms of sample efficiency' while maintaining computational efficiency, using batch normalization as a core component. The original paper's claim of novelty for combining batch normalization with distributional cross-entropy loss in an actor-critic framework is challenged by CrossQ[65]'s prior demonstration that batch normalization alone (with different architectural choices) can achieve similar sample efficiency goals.

Evidence

Evidence 1 - **Rationale**: Both papers claim to introduce novel algorithms achieving state-of-the-art sample efficiency in continuous control by extending actor-critic methods with architectural innovations including batch normalization. CrossQ[65] demonstrates this was achieved earlier. - **Original**: we introduce xqc : a well-motivated, sample-efficient deep actor-critic algorithm built upon soft actor-critic that embodies these optimization-aware principles. we achieve state-of-the-art sample efficiency across 55 proprioception and 15 vision-based continuous control tasks - **Candidate**: we introduce crossq: a lightweight algorithm for continuous control tasks that makes careful use of batch normalization and removes target networks to surpass the current state-of-the-art in sample efficiency while maintaining a low utd ratio of 1.

Evidence 2 - **Rationale**: Both papers present their algorithms as simple, efficient extensions to SAC that achieve state-of-the-art sample efficiency. CrossQ[65]'s prior work on batch normalization-based improvements to actor-critic methods challenges the novelty of the original paper's approach. - **Original**: xqc , a simple and efficient extension to soft actor critic, uses the powerful synergy between bn, wn, and a distributional critic with a ce bellman error loss for sample-efficient learning. - **Candidate**: crossq's contributions are threefold: (1) it matches or surpasses current state-of-the-art methods in terms of sample efficiency, (2) it substantially reduces the computational cost compared to redq and droq, (3) it is easy to implement, requiring just a few lines of code on top of sac.

Evidence 3 - **Rationale**: Both papers use batch normalization as a core architectural component for improving sample efficiency in actor-critic algorithms. CrossQ[65] demonstrates this approach was established earlier, challenging the novelty of incorporating BN into the critic architecture. - **Original**: batch normalization. xqc uses bn layers directly on the network input and after each linear layer (figure 5). following bhatt et al. (2024), we implement a joined forward pass to automatically calculate the bn running statistics on the joined (s, a) and (s′, a′) distribution (bhatt et al., 2024), to... - **Candidate**: we introduce crossq: a lightweight algorithm for continuous control tasks that makes careful use of batch normalization and removes target networks to surpass the current state-of-the-art in sample efficiency

8. CTRL-B: Back-End-Of-Line Configuration Pathfinding Using Cross-Technology Transferable Reinforcement Learning

URL: View paper

Brief Assessment

CTRL-B[63] focuses on back-end-of-line semiconductor configuration optimization using policy gradient methods for chip design, not on general continuous control tasks with actor-critic architectures.

9. Price-taker Bidding and Pricing Strategy Using Deep Deterministic Policy Gradient Algorithm with Transformer Neural Networks

URL: View paper

Brief Assessment

Price-Taker Transformer[67] focuses on electricity market bidding using DDPG with transformer networks for a specific domain application, not on general sample-efficient deep RL with normalization techniques and distributional critics.

10. Hyperspherical Normalization for Scalable Deep Reinforcement Learning

URL: View paper

Prior Art Analysis

Hyperspherical Normalization[68] demonstrates that prior work exists combining normalization techniques with distributional critics for sample-efficient deep RL. Both papers propose actor-critic algorithms extending soft actor-critic with normalization strategies and distributional value estimation to improve sample efficiency. The candidate paper (SimbaV2) explicitly uses hyperspherical normalization with distributional value estimation, achieving state-of-the-art performance on continuous control tasks, which overlaps substantially with the original paper's core contribution of combining batch normalization, weight normalization, and distributional cross-entropy loss.

Evidence

Evidence 1 - **Rationale**: Both papers propose extensions to soft actor-critic that combine normalization techniques with distributional value estimation to achieve state-of-the-art sample efficiency on continuous control tasks. This demonstrates that the combination of these techniques was not novel to the original paper. - **Original**: we introduce xqc : a well-motivated, sample-efficient deep actor-critic algorithm built upon soft actor-critic that embodies these optimization-aware principles. we achieve state-of-the-art sample efficiency across 55 proprioception and 15 vision-based continuous control tasks - **Candidate**: we introduce simbav2, a novel rl architecture designed to stabilize optimization by (i) constraining the growth of weight and feature norm by hyperspherical normalization; and (ii) using a distributional value estimation with reward scaling to maintain stable gradients under varying reward magnitude...

Evidence 2 - **Rationale**: The candidate paper's combination of normalization (hyperspherical normalization constrains weight and feature norms) with distributional value estimation directly parallels the original paper's claimed novel combination of batch normalization, weight normalization, and distributional critic. - **Original**: xqc , a simple and efficient extension to soft actor critic, uses the powerful synergy between bn, wn, and a distributional critic with a ce bellman error loss for sample-efficient learning. - **Candidate**: we introduce simbav2, a novel rl architecture designed to stabilize optimization by (i) constraining the growth of weight and feature norm by hyperspherical normalization; and (ii) using a distributional value estimation with reward scaling to maintain stable gradients

Contribution 2: Hessian eigenvalue analysis of deep RL critic optimization

Description: The authors conduct a systematic eigenvalue analysis of the critic's Hessian to investigate how architectural components affect the optimization landscape. They demonstrate that distributional critics with cross-entropy loss produce condition numbers orders of magnitude smaller than mean squared error losses, providing a principled explanation for performance differences.

This contribution was assessed against **1 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Spectral-Risk Multi-Objective Reinforcement Learning

URL: View paper

Brief Assessment

Spectral-Risk Multi-Objective[51] focuses on multi-objective RL with spectral risk measures and does not conduct Hessian eigenvalue analysis of critic optimization landscapes or compare distributional versus MSE losses.

Contribution 3: Theoretical analysis of cross-entropy versus squared error loss conditioning

Description: The authors provide formal analysis showing that cross-entropy loss has bounded gradients and upper-bounded condition numbers, while mean squared error loss has unbounded gradients and cannot upper-bound the condition number. This theoretical framework explains why cross-entropy loss creates better-conditioned optimization landscapes for deep reinforcement 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. Scaling laws from the data manifold dimension

URL: View paper

Brief Assessment

Data Manifold Dimension[59] focuses on scaling laws from data manifold dimension in supervised learning and language modeling, not on optimization conditioning in deep reinforcement learning. The paper analyzes how loss scales with model parameters based on intrinsic dimensionality, which is a fundamentally different theoretical framework than analyzing Hessian conditioning for RL critic networks.

2. Stop regressing: Training value functions via classification for scalable deep rl

URL: View paper

Brief Assessment

Stop Regressing Classification[54] focuses on empirical scalability improvements from using cross-entropy for value functions across various domains, but does not provide the formal theoretical analysis of gradient bounds and condition numbers that characterizes the original paper's contribution.

3. Classification-Based Q-Value Estimation for Continuous Actor-Critic Reinforcement Learning

URL: View paper

Brief Assessment

Classification-Based Q-Value[57] focuses on reformulating Q-value estimation as classification using CE/KL losses for continuous control, emphasizing empirical stability and overestimation reduction. It does not provide formal theoretical analysis of gradient bounds, condition numbers, or Hessian eigenspectra that characterize optimization landscape conditioning.

4. Conditioned reinforcement learning for few-shot imitation

URL: View paper

Brief Assessment

Conditioned Few-Shot Imitation[60] focuses on demonstration-conditioned reinforcement learning for few-shot imitation tasks, not on theoretical analysis of loss function conditioning properties in deep RL optimization landscapes.

5. Deep Belief Markov Models for POMDP Inference

URL: View paper

Brief Assessment

Deep Belief Markov[55] focuses on POMDP inference using deep belief networks for state estimation, not on analyzing loss function conditioning properties in deep RL optimization landscapes.

6. The central role of the loss function in reinforcement learning

URL: View paper

Brief Assessment

Central Loss Function[52] focuses on theoretical sample efficiency bounds and cost-sensitive classification, not on optimization landscape conditioning (Hessian eigenspectra, condition numbers) in deep RL as analyzed in the original paper.

7. Rectifying Regression in Reinforcement Learning

URL: View paper

Brief Assessment

Rectifying Regression[58] focuses on prediction objectives (mean absolute error vs mean squared error) and their alignment with loss functions for controlling suboptimality gaps in linear RL. The original paper analyzes optimization landscape conditioning through Hessian eigenspectra, gradient bounds, and condition numbers in deep RL with function approximation—a fundamentally different theoretical framework.

8. State-aware perturbation optimization for robust deep reinforcement learning

URL: View paper

Brief Assessment

State-Aware Perturbation[53] focuses on adversarial robustness through perturbation optimization and mentions cross-entropy and mean squared error losses only in passing as loss functions for different methods. It does not provide theoretical analysis of loss conditioning properties or their impact on optimization landscapes in deep RL.

9. Is Value Functions Estimation with Classification Plug-and-play for Offline Reinforcement Learning?

URL: View paper

Brief Assessment

Classification Plug-and-Play[56] focuses on empirical benchmarking of cross-entropy versus MSE in offline RL settings without providing theoretical analysis of conditioning properties, gradient bounds, or condition numbers that characterize the optimization landscape.

10. Real-Time Adaptive Loss Functions for Generative Models Using Reinforcement Learning and Meta-Learning

URL: View paper

Brief Assessment

Adaptive Loss Functions[61] focuses on dynamically adapting loss functions using RL and meta-learning for generative models, not on theoretical analysis of cross-entropy versus MSE conditioning in deep RL optimization landscapes.

Appendix: Text Similarity Detection

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

References

- [0] XQC: Well-conditioned Optimization Accelerates Deep Reinforcement Learning View paper
- [1] On the data-efficiency with contrastive image transformation in reinforcement learning View paper
- [2] Efficient offline reinforcement learning: The critic is critical View paper
- [3] Boosting On-Policy Actorâ□□Critic With Shallow Updates in Critic View paper
- [4] Data-efficient deep reinforcement learning for dexterous manipulation View paper
- [5] Data valuation using reinforcement learning View paper
- [6] Simplicial embeddings improve sample efficiency in actor-critic agents View paper
- [7] Improving sample-efficiency of model-free reinforcement learning algorithms on image inputs with representation learning View paper
- [8] Policy Correction and State-Conditioned Action Evaluation for Few-Shot Lifelong Deep Reinforcement Learning View paper
- [9] Adaptive critic learning for approximate optimal event-triggered tracking control of nonlinear systems with prescribed performances View paper
- [10] Data-efficient controller tuning and reinforcement learning View paper
- [11] Rethinking value function learning for generalization in reinforcement learning View paper
- [12] CTD4 - A Deep Continuous Distributional Actor-Critic Agent with a Kalman Fusion of Multiple Critics View paper
- [13] Efficient deep reinforcement learning-enabled recommendation View paper
- [14] Is a good representation sufficient for sample efficient reinforcement learning? View paper
- [15] Iterated Q-Network: Beyond One-Step Bellman Updates in Deep Reinforcement Learning View paper
- [16] Deep Reinforcement Learning Algorithms for Profitable Stock Trading Strategies View paper
- [17] Learning and Reusing Primitive Behaviours to Improve Hindsight Experience Replay Sample Efficiency View paper
- [18] Constraint-conditioned policy optimization for versatile safe reinforcement learning View paper
- [19] Learning to learn: Meta-critic networks for sample efficient learning View paper
- [20] Guided cooperation in hierarchical reinforcement learning via model-based rollout View paper
- [21] Investigating effects of centralized learning decentralized execution on team coordination in the level based foraging environment as a sequential social dilemma View paper
- [22] Deep Reinforcement Learning Based on Parked Vehicles-Assisted for Task Offloading in Vehicle Edge Computing View paper
- [23] Value-consistent representation learning for data-efficient reinforcement learning View paper
- [24] Interpolated Policy Gradient: Merging On-Policy and Off-Policy Gradient Estimation for Deep Reinforcement Learning View paper
- [25] Investigating the multi-objective optimization of quality and efficiency using deep reinforcement learning View paper
- [26] Recommendation of deep reinforcement learning based on value function considering error reduction View paper
- [27] Robotic Offline RL from Internet Videos via Value-Function Learning View paper
- [28] Memory-efficient Reinforcement Learning with Value-based Knowledge Consolidation View paper
- [29] Advancing data-efficiency in reinforcement learning View paper
- [30] Towards Data-efficient AI: Theoretical analysis and experimental validation of new exploration algorithms for Reinforcement Learning View paper
- [31] Sample-efficient deep reinforcement learning for control, exploration and safety View paper
- [32] Efficient Data Usage for Planning and Reinforcement Learning View paper
- [33] Investigating Sample Efficient Deep Reinforcement Learning View paper
- [34] Robot Path Planning in Crowd Environments via Dynamic Interaction Modeling and Deep Reinforcement Learning View paper
- [35] Bridging Exploration and General Function Approximation in Reinforcement Learning: Provably Efficient Kernel and Neural Value Iterations View paper
- [36] A novel multi-step Q-learning method to improve data efficiency for deep reinforcement learning View paper
- [37] Value-Based Deep RL Scales Predictably View paper
- [38] Robotic Offline RL from Internet Videos via Value-Function Pre-Training View paper
- [39] Interaction-Efficient Reinforcement Learning: Matching the Real World Data Availability View paper
- [40] Value Function Initialization for Knowledge Transfer and Jump-start in Deep Reinforcement Learning View paper
- [41] Efficient RL Training for Reasoning Models via Length-Aware Optimization View paper
- [42] Compute-Optimal Scaling for Value-Based Deep RL View paper
- [43] A nonlinear robust control method based on distributed deep reinforcement learning View paper
- [44] Adaptive Exploration for Data-Efficient General Value Function Evaluations View paper
- [45] Satisficing Exploration for Deep Reinforcement Learning View paper
- [46] Exploration of DeepSeek System Supported by Deep Reinforcement Learning in Architectural Creative Design View paper
- [47] AM-PPO:(Advantage) Alpha-Modulation with Proximal Policy Optimization View paper
- [48] Data Efficient Deep Reinforcement Learning With Action-Ranked Temporal Difference Learning View paper
- [49] A Feedback-based Decision-Making Mechanism for Actor-Critic Deep Reinforcement Learning View paper
- [50] ViVa: Video-Trained Value Functions for Guiding Online RL from Diverse Data View paper
- [51] Spectral-Risk Multi-Objective Reinforcement Learning View paper
- [52] The central role of the loss function in reinforcement learning View paper
- [53] State-aware perturbation optimization for robust deep reinforcement learning View paper
- [54] Stop regressing: Training value functions via classification for scalable deep rl View paper
- [55] Deep Belief Markov Models for POMDP Inference View paper
- [56] Is Value Functions Estimation with Classification Plug-and-play for Offline Reinforcement Learning? View paper
- [57] Classification-Based Q-Value Estimation for Continuous Actor-Critic Reinforcement Learning View paper
- [58] Rectifying Regression in Reinforcement Learning View paper
- [59] Scaling laws from the data manifold dimension View paper
- [60] Conditioned reinforcement learning for few-shot imitation View paper
- [61] Real-Time Adaptive Loss Functions for Generative Models Using Reinforcement Learning and Meta-Learning View paper
- [62] Combining policy gradient and Q-learning View paper

- [63] CTRL-B: Back-End-Of-Line Configuration Pathfinding Using Cross-Technology Transferable Reinforcement Learning View paper
- [64] Striving for simplicity and performance in off-policy DRL: Output normalization and non-uniform sampling View paper
- [65] Crossq: Batch normalization in deep reinforcement learning for greater sample efficiency and simplicity View paper
- [66] Autonomous Navigation of Mobile Robots in Complex Environments with Global Path Smoothing and Adaptive Local Control View paper
- [67] Price-taker Bidding and Pricing Strategy Using Deep Deterministic Policy Gradient Algorithm with Transformer Neural Networks View paper
- [68] Hyperspherical Normalization for Scalable Deep Reinforcement Learning View paper
- [69] Personalized Recommendation System Based on Deep Reinforcement Learning View paper
- [70] Optimistic Actor-Critic with Parametric Policies: Unifying Sample Efficiency and Practicality View paper
- [71] Optimizing Game Strategies with Deep Reinforcement Learning: A Framework for Intelligent Decision-Making View paper