Novelty Assessment Report

Paper: Sampling Complexity of TD and PPO in RKHS
PDF URL: https://openreview.net/pdf?id=5gUMhTUDi0
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2026-01-07

Abstract

We revisit Proximal Policy Optimization (PPO) from a function-space perspective. Our analysis decouples policy evaluation and improvement in a reproducing kernel Hilbert space (RKHS): (i) A kernelized temporal-difference (TD) critic performs efficient RKHS-gradient updates using only one-step state–action transition samples. (ii) a KL-regularized, natural-gradient policy step exponentiates the evaluated action-value, recovering a PPO/TRPO-style proximal update in continuous state-action spaces. We provide non-asymptotic, instance-adaptive guarantees whose rates depend on RKHS entropy, unifying tabular, linear, Sobolev, Gaussian, and Neural Tangent Kernel (NTK) regimes, and we derive a sampling rule for the proximal update that ensures the optimal $k^{-1/2}$ convergence rate for stochastic optimization. Empirically, the theory-aligned schedule improves stability and sample efficiency on common control tasks (e.g., CartPole, Acrobot), while our TD-based critic attains favorable throughput versus a GAE baseline. Altogether, our results place PPO on a firmer theoretical footing beyond finite-dimensional assumptions and clarify when RKHS-proximal updates with kernel-TD critics yield global policy improvement with practical efficiency.

Core Task Landscape

This paper addresses: **Sampling Complexity of Temporal Difference and Proximal Policy Optimization in Reproducing Kernel Hilbert Spaces**

A total of **24 papers** were analyzed and organized into a taxonomy with **15 categories**.

Taxonomy Overview

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

- **Temporal Difference Learning in RKHS**
- **Proximal Policy Optimization in RKHS and Function Spaces**
- **Distributional and Regularized Policy Iteration**

Complete Taxonomy Tree

- Sampling Complexity of Temporal Difference and Proximal Policy Optimization in Reproducing Kernel Hilbert Spaces Survey Taxonomy
- Temporal Difference Learning in RKHS
  - Kernel-Based Least Squares Temporal Difference
  - Regularized LSTD with Convergence Guarantees (2 papers)
    - [5] A non-asymptotic analysis of non-parametric temporal-difference learning (Berthier, 2022) View paper
    - [6] Optimal policy evaluation using kernel-based temporal difference methods (Duan, 2024) View paper
  - Gradient-Corrected and Efficient LSTD Variants (2 papers)
    - [10] Kernel-based least squares temporal difference with gradient correction (Tianheng Song, 2015) View paper
    - [16] An efficient L2-norm regularized least-squares temporal difference learning algorithm (Sheng-Lei Chen, 2013) View paper
  - Gradient Temporal Difference in RKHS
  - Kernelized GTD with Dimensionality Breaking (2 papers)
    - [18] Breaking bellman's curse of dimensionality: Efficient kernel gradient temporal difference (Koppel, 2017) View paper
    - [23] Policy Evaluation in Continuous MDPs with Efficient Kernelized Gradient Temporal Difference (Koppel, 2017) View paper
  - Multi-Agent Kernelized Gradient TD (2 papers)
    - [7] Kernel-based decentralized policy evaluation for reinforcement learning (Jiamin Liu, 2024) View paper
    - [11] Finite-sample analysis of multi-agent policy evaluation with kernelized gradient temporal difference (Paulo Heredia, 2020) View paper
  - Online and Sparse Kernel TD Methods
  - Online Selective Kernel TD with Sparsification (2 papers)
    - [3] Online Attentive Kernel-Based Temporal Difference Learning (Guang Yang, 2023) View paper
    - [9] Online selective kernel-based temporal difference learning (Xingguo Chen, 2013) View paper
  - Sparse Kernel LSTD Variants (1 papers)
    - [17] A sparse kernel-based least-squares temporal difference algorithm for reinforcement learning (Xin Xu, 2006) View paper
  - Gaussian Process and Scalable Kernel TD (1 papers)
  - [1] Gaussian process temporal-difference learning with scalability and worst-case performance guarantees (Qin Lu, 2021) View paper
  - Application-Specific Kernel TD (5 papers)
  - [8] Kernel temporal difference based reinforcement learning for brain machine interfaces (Xiang Shen, 2021) View paper
  - [12] Kernel temporal differences for neural decoding (Jihye Bae, 2015) View paper
  - [15] Kernel temporal differences for reinforcement learning with applications to brain machine interfaces (Jihye Bae, 2013) View paper

- [19] Kernel Temporal Differences for EEG-based Reinforcement Learning Brain Machine Interfaces. (Bhoj Raj Thapa, 2022) View paper
- [22] Reinforcement learning via kernel temporal difference. (Jihye Bae, 2012) View paper

- Proximal Policy Optimization in RKHS and Function Spaces
  - RKHS-Based PPO with Sampling Complexity Analysis ★ (1 papers)
  - [0] Sampling Complexity of TD and PPO in RKHS (Anon et al., 2026) View paper
  - Kernel Policy Networks for Stability (1 papers)
  - [2] Residual kernel policy network: Enhancing stability and robustness in rkhs-based reinforcement learning (Y Zhang, 2025) View paper
  - Correntropy-Induced Metric PPO Variants (2 papers)
  - [20] CIM-PPO:Proximal Policy Optimization with Liu-Correntropy Induced Metric (Guo, 2022) View paper
  - [24] PPO-CIM: Proximal Policy Optimization with Correntropy Induced Metric (Y Guoa, n.d.) View paper
  - PPO with RKHS-Based Information State (1 papers)
  - [21] Reinforcement learning in partially observable environments using approximate information state (Amit Kumar Sinha, 2021) View paper
  - PPO in Diverse Data and Specialized Settings (1 papers)
  - [13] Towards Specialized Reinforcement Learning From Diverse Data (Chang, 2024) View paper
- Distributional and Regularized Policy Iteration
  - Distributional TD in Hilbert Spaces (1 papers)
  - [4] Statistical Efficiency of Distributional Temporal Difference Learning and Freedman's Inequality in Hilbert Spaces (Peng Yang, 2024) View paper
  - Regularized Policy Iteration with Non-Parametric Approximation (1 papers)
  - [14] Regularized policy iteration (Mohammad Ghavamzadeh, 2008) View paper

Narrative

Core task: sampling complexity of temporal difference and proximal policy optimization in reproducing kernel Hilbert spaces. The field structure reflects a natural division into three main branches. The first branch, Temporal Difference Learning in RKHS, encompasses a substantial body of work on kernel-based value function approximation, ranging from foundational methods like Gaussian Process TD Learning[1] and early sparse approaches such as Sparse Kernel LSTD[17] to more recent algorithmic refinements including Online Attentive Kernel TD[3] and Optimal Kernel TD[6]. The second branch, Proximal Policy Optimization in RKHS and Function Spaces, explores policy optimization in infinite-dimensional settings, with works like Residual Kernel Policy Network[2] and variants employing alternative loss metrics such as PPO Liu Correntropy[20]. The third branch, Distributional and Regularized Policy Iteration, addresses regularization strategies and distributional perspectives, exemplified by Regularized Policy Iteration[14] and Distributional TD Freedman Inequality[4]. Together, these branches capture the interplay between function approximation theory, sample efficiency, and algorithmic design in kernel-based reinforcement learning.

A particularly active line of work centers on establishing finite-sample guarantees for kernel TD methods, with studies like Nonparametric TD Analysis[5] and Optimal Kernel TD[6] investigating convergence rates and sample complexity under various kernel choices. In contrast, the PPO-focused branch remains less densely populated, with only a handful of papers exploring how proximal updates behave in RKHS settings. Sampling Complexity TD PPO[0] sits at the intersection of these themes, directly addressing sample efficiency for both TD and PPO in reproducing kernel Hilbert spaces. Compared to neighboring works such as Residual Kernel Policy Network[2], which emphasizes architectural design, or Nonparametric TD Analysis[5], which focuses on TD convergence alone, the original paper[0] provides a unified treatment of sampling complexity across both value-based and policy-gradient paradigms. This dual focus highlights ongoing questions about how kernel smoothness, exploration strategies, and policy parameterization jointly influence learning efficiency in continuous or large state spaces.

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 theoretical sampling complexity and convergence guarantees for PPO in RKHS, while siblings explore practical modifications and applications. Siblings include metric-based variants (correntropy), architectural enhancements for stability, specialized application settings, and partial observability solutions using RKHS embeddings. The original leaf is distinguished by its emphasis on formal sample efficiency analysis rather than algorithmic modifications or domain-specific applications.

Similarities: - All subtopics operate within or relate to the RKHS framework for policy optimization - All involve PPO or policy gradient methods as the core algorithmic foundation - Multiple subtopics leverage RKHS properties for theoretical or practical improvements to reinforcement learning

Differences: - Original leaf: Theoretical focus on sampling complexity bounds and convergence rates; Siblings: Algorithmic modifications (metrics, architectures) or application-specific adaptations - Original leaf: Analyzes sample efficiency guarantees; Correntropy sibling: Replaces divergence measures with alternative metrics - Original leaf: Fully observable settings with theory; RKHS-based information state sibling: Addresses partial observability through state embeddings - Kernel policy networks sibling: Emphasizes architectural stability; Original leaf: Emphasizes theoretical performance guarantees

Suggested Search Directions: - Investigate whether correntropy-metric PPO variants have developed their own sampling complexity theory - Explore connections between RKHS information state embeddings and sample efficiency in partially observable domains - Examine if kernel policy network architectures can be analyzed through the sampling complexity lens of the original leaf

Sibling Subtopics

- **Correntropy-Induced Metric PPO Variants** (leaves: 1, papers: 2)
- Scope: PPO algorithms replacing KL divergence with correntropy-induced metrics in reproducing kernel Hilbert spaces.
- Exclude: Standard KL-based PPO and RKHS methods without metric modifications are excluded; they belong to other leaves.
- **Kernel Policy Networks for Stability** (leaves: 1, papers: 1)
- Scope: Policy network architectures using kernel methods in RKHS to enhance stability and robustness.
- Exclude: Theoretical sampling analysis and metric-based PPO are excluded; they belong to sibling leaves.
- **PPO in Diverse Data and Specialized Settings** (leaves: 1, papers: 1)
- Scope: PPO applications addressing specialized learning scenarios with diverse data sources and complexity measures.
- Exclude: Standard PPO formulations and RKHS-specific theoretical analysis are excluded; they belong to other leaves.
- **PPO with RKHS-Based Information State** (leaves: 1, papers: 1)
- Scope: PPO methods using approximate information states constructed via RKHS embeddings for partial observability.

- Exclude: Fully observable settings and non-RKHS information representations are excluded; they belong to other categories.

Contributions Analysis

Overall novelty summary. The paper contributes a unified RKHS framework for PPO that decouples policy evaluation (via kernelized TD) and improvement (via natural-gradient proximal updates), accompanied by non-asymptotic sampling complexity guarantees. It resides in the 'RKHS-Based PPO with Sampling Complexity Analysis' leaf, which contains only this single paper. This positioning reflects a sparse research direction: while the broader 'Proximal Policy Optimization in RKHS and Function Spaces' branch includes five leaves, most sibling leaves address architectural stability or metric modifications rather than sampling complexity theory.

The taxonomy reveals that neighboring work clusters around two distinct themes. The 'Temporal Difference Learning in RKHS' branch (comprising four subtopics and roughly fifteen papers) focuses on kernel-based value approximation—methods like Gaussian Process TD and sparse LSTD variants—but does not integrate policy optimization. Meanwhile, sibling leaves in the PPO branch (e.g., 'Kernel Policy Networks for Stability', 'Correntropy-Induced Metric PPO Variants') emphasize empirical robustness or alternative divergence measures without theoretical sampling analysis. The original paper bridges these areas by combining kernel TD evaluation with proximal policy steps under a single convergence framework.

Among twenty-three candidates examined, the kernelized TD critic (Contribution 1) encountered three potentially overlapping works out of eight reviewed, while the KL-regularized natural-gradient update (Contribution 2) found one refutable candidate among six. The instance-adaptive convergence guarantees (Contribution 3) showed no clear refutation across nine candidates. These statistics suggest that the TD component has more substantial prior work in kernel methods, whereas the unified sampling complexity analysis across tabular, linear, Sobolev, and NTK regimes appears less directly anticipated by the limited search scope.

Given the restricted search scale (top-23 semantic matches), the analysis captures immediate neighbors but cannot claim exhaustive coverage. The sparse taxonomy leaf and the absence of refutation for the unified convergence guarantees hint at novelty in the theoretical synthesis, though the kernelized TD and natural-gradient components build on established kernel RL techniques. A broader literature review would clarify whether the integration of these elements under a single sampling complexity framework represents a significant theoretical advance or an incremental unification of known results.

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

Contribution 1: Kernelized temporal-difference critic with RKHS-gradient updates

Description: The authors introduce a kernel gradient-based TD evaluator in a reproducing kernel Hilbert space that acts as an implicit preconditioner and achieves geometric convergence without requiring costly matrix inversions. The evaluator leverages N-step TD learning and provides non-asymptotic TD-error bounds attaining the minimax rate up to logarithmic factors.

This contribution was assessed against **8 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. A non-asymptotic analysis of non-parametric temporal-difference learning

URL: View paper

Prior Art Analysis

Nonparametric TD Analysis[5] demonstrates that kernelized temporal-difference learning in RKHS was already studied prior to the original paper's submission. The candidate paper presents a comprehensive non-asymptotic analysis of non-parametric TD(0) in reproducing kernel Hilbert spaces, including convergence guarantees and explicit rates. Both papers formulate TD updates as RKHS-gradient iterations and provide theoretical convergence analysis, though with different technical approaches (the candidate uses regularized fixed-point equations while the original uses implicit preconditioning).

Evidence

Evidence 1 - **Rationale**: Both papers address kernelized TD in RKHS with non-asymptotic convergence guarantees, establishing that this approach was already explored before the original paper's claimed contribution. - **Original**: we introduce a kernel, gradient-based td evaluator in an rkhs that acts as an implicit preconditioner and attains geometric convergence without costly matrix inversions. the evaluator leverages n-step td learning for any n≥1 and provides non-asymptotic td-error bounds that attain the minimax rate (u… - **Candidate**: temporal-difference learning is a popular algorithm for policy evaluation. in this paper, we study the convergence of the regularized non-parametric td(0) algorithm, in both the independent and markovian observation settings. in particular, when td is performed in a universal reproducing kernel hilb...

Evidence 2 - **Rationale**: Both papers provide explicit non-asymptotic convergence rates for kernelized TD, with the candidate paper establishing this result earlier, thereby challenging the novelty claim of being the first to provide such guarantees. - **Original**: we provide non-asymptotic, instance-adaptive guarantees whose rates depend on rkhs entropy, unifying tabular, linear, sobolev, gaussian, and neural tangent kernel (ntk) regimes - **Candidate**: we provide explicit convergence rates that depend on a source condition relating the regularity of the optimal value function to the rkhs. we illustrate this convergence numerically on a simple continuous-state markov reward process.

---

2. Finite-sample analysis of multi-agent policy evaluation with kernelized gradient temporal difference

URL: View paper

Brief Assessment

Multiagent Kernelized GTD[11] focuses on distributed multi-agent policy evaluation with consensus protocols under time-varying networks, whereas the original paper proposes a centralized kernel gradient-based TD evaluator for single-agent PPO with geometric convergence guarantees and minimax-optimal rates.

---

3. Kernel-based decentralized policy evaluation for reinforcement learning

URL: View paper

Brief Assessment

Decentralized Kernel Policy Evaluation[7] focuses on decentralized multi-agent policy evaluation using regression-based methods in RKHS, whereas the original paper proposes a centralized kernel TD critic with specific RKHS-gradient updates for policy optimization. The candidate's regression-based approach and decentralized setting differ fundamentally from the original's centralized kernel TD framework.

---

4. Gain Function Tracking in the Feedback Particle Filter

URL: View paper

Brief Assessment

Gain Function Tracking[25] focuses on gain function approximation for the feedback particle filter using RKHS-based differential TD-learning, not on policy evaluation in reinforcement learning with kernelized TD critics and geometric convergence guarantees.

---

5. Policy Evaluation in Continuous MDPs with Efficient Kernelized Gradient Temporal Difference

URL: View paper

Prior Art Analysis

Efficient Kernelized GTD[23] demonstrates that prior work exists on kernelized temporal-difference methods with RKHS-gradient updates for policy evaluation. The candidate paper presents a kernel gradient-based TD evaluator in a reproducing kernel Hilbert space that performs efficient gradient updates and achieves geometric convergence without costly matrix inversions. This directly addresses the same technical problem as the original contribution, using similar mathematical frameworks (RKHS, kernel methods, temporal difference learning) and achieving comparable goals (efficient policy evaluation without matrix inversions).

Evidence

Evidence 1 - **Rationale**: Both papers propose kernel-based TD methods in RKHS that avoid costly matrix inversions. The candidate explicitly describes generalizing stochastic quasi-gradient methods to RKHS for memory-efficient policy evaluation. - **Original**: we introduce a kernel, gradient-based td evaluator in an rkhs that acts as an implicit preconditioner and attains geometric convergence without costly matrix inversions. - **Candidate**: we propose to tackle this memory bottleneck by requiring memory efficiency in both the function sample path and in its limit, whose complexity in the worst case is defined by the metric entropy of the state space (corollary 1). to find a memory-efficient sample path in the function space, we generalize ...

Evidence 2 - **Rationale**: Both papers describe RKHS-gradient updates for TD learning using state-action transition samples. The candidate's equation (10) shows the functional gradient in RKHS using temporal differences. - **Original**: a kernelized temporal-difference (td) critic performs efficient rkhs-gradient updates using only one-step state-action transition samples. - **Candidate**: to perform stochastic descent in function space h, we need a stochastic approximate of (9) evaluated at a state-action-state triple (x,π(x),y), which together with the regularizer yields ∇vj(v,δ; x,π(x),y) (10) =[γκ(y,·)-κ(x,·)][r(x,π(x),y)+γv(y)-v(x)]+λv where δ := r(x,π(x),y) + γv(y) -v(x) is defi...

Evidence 3 - **Rationale**: Both papers provide convergence guarantees for kernelized TD methods in RKHS. The candidate establishes almost sure convergence to the Bellman fixed point, while the original claims minimax-rate TD-error bounds. - **Original**: the evaluator leverages n-step td learning for any n≥1 and provides non-asymptotic td-error bounds that attain the minimax rate (up to logarithmic factors). - **Candidate**: our main result is a memory-efficient, non-parametric, stochastic method that converges to the bellman fixed point almost surely when it belongs to a reproducing kernel hilbert space (rkhs). the hypothesis that the value function belongs to a rkhs restricts the relationship between rewards and value t...

---

6. RKHS Temporal Difference Learning

URL: View paper

Prior Art Analysis

RKHS TD Learning[26] demonstrates that kernelized temporal-difference learning with RKHS-gradient updates was previously proposed and implemented. The candidate paper presents a kernel gradient-based TD evaluator in RKHS that performs gradient updates using kernel functions, which directly overlaps with the original paper's claimed contribution. Both papers formulate TD learning in RKHS using kernel gradient descent methods, derive update rules based on RKHS inner products, and analyze convergence properties. The candidate paper explicitly introduces 'a reproducing kernel hilbert space version of sarsa(λ)' with kernel-based gradient updates, predating the original paper's claim of novelty.

Evidence

Evidence 1 - **Rationale**: Both papers introduce kernelized TD learning in RKHS. The candidate explicitly presents this as a novel contribution, demonstrating prior work exists. - **Original**: we introduce a kernel, gradient-based td evaluator in an rkhs that acts as an implicit preconditioner and attains geometric convergence without costly matrix inversions. the evaluator leverages n-step td learning for any n≥1 and provides non-asymptotic td-error bounds - **Candidate**: we introduce a reproducing kernel hilbert space version of sarsa(λ) including a simple and intuitive formula for the eligibility trace... we introduce a memory efficient version which dramatically reduces the number of terms in the representation of the estimated value function

Evidence 2 - **Rationale**: Both papers derive kernel-based gradient update rules for TD learning in RKHS using similar mathematical formulations with kernel functions k(·,·). - **Original**: inspired by kernel gradient descent (ding et al., 2024; lin & zhou, 2018; raskutti et al., 2014), we propose the following updating rule in the rkhs: ft+1 = (1-α t)ft -η t nx i=1 ft(ω(i) 0)-r(ω (i) 0)-γf t(ω(i) 1) k(ω(i) 0 ,·) - **Candidate**: by substituting (10) into (8), we get wt+1 = wt -ηterr(st,at,rt) t∑ i=t0 (γλ)t-iφ(si,ai) -ξwt... and assuming that w0 = 0 we see that wt = ∑t-1 i=1 αik((si,ai),·) which leads us to the dual update formulation

---

7. Breaking bellman's curse of dimensionality: Efficient kernel gradient temporal difference

URL: View paper

Brief Assessment

Breaking Bellman Curse[18] focuses on memory-efficient sparse subspace projections for policy evaluation in infinite spaces, while the original paper introduces a kernel gradient-based TD evaluator with N-step learning and non-asymptotic bounds for policy optimization within PPO framework.

---

8. A sparse kernel-based least-squares temporal difference algorithm for reinforcement learning

URL: View paper

Brief Assessment

Sparse Kernel LSTD[17] focuses on kernel-based LSTD (least-squares temporal difference) methods for tabular settings, whereas the original paper introduces a gradient-based kernel TD critic that avoids matrix inversions and provides non-asymptotic bounds across multiple RKHS regimes (tabular, Sobolev, NTK, Gaussian).

Contribution 2: KL-regularized natural-gradient policy update for continuous spaces

Description: The authors design a KL-regularized proximal update implementable in continuous action spaces that exponentiates the value estimate. They explicitly quantify the per-iteration sample size needed to achieve intended improvement, addressing a gap where policy expectations are often treated as exact or left unspecified.

This contribution was assessed against **6 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. FCPortugal-multi-robot action learning

URL: View paper

Brief Assessment

FCPortugal Multi Robot[31] focuses on multi-robot coordination for set plays in simulated robotic soccer using TD3 algorithm, not on KL-regularized natural-gradient policy updates in continuous action spaces.

---

2. SACrificing Intuition

URL: View paper

Brief Assessment

SACrificing Intuition[32] focuses on replacing SAC's entropy bonus with KL regularization against a reference policy, not on implementing KL-regularized proximal updates in continuous action spaces with explicit sample-size quantification for policy improvement as described in the original contribution.

---

3. Achieving Zero Constraint Violation for Constrained Reinforcement Learning via Conservative Natural Policy Gradient Primal-Dual Algorithm

URL: View paper

Brief Assessment

Zero Constraint Violation CNPG[28] focuses on constrained RL with zero constraint violation guarantees, not on KL-regularized proximal updates for continuous action spaces or sample complexity quantification for policy expectations.

---

4. Optimal Rates of Convergence for Entropy Regularization in Discounted Markov Decision Processes

URL: View paper

Brief Assessment

Entropy Regularization Convergence Rates[27] focuses on theoretical convergence rates of entropy regularization in discrete MDPs, not on implementable policy updates in continuous action spaces or sample complexity quantification for policy improvement steps.

---

5. Compatible natural gradient policy search

URL: View paper

Prior Art Analysis

Compatible Natural Gradient[29] demonstrates prior work on KL-regularized policy updates in continuous action spaces. The candidate paper presents a trust-region optimization framework with KL-divergence constraints that yields closed-form policy updates for continuous actions, specifically showing that the natural gradient is equivalent to trust-region optimization when using natural parameterization. The paper explicitly derives update rules for continuous Gaussian policies with KL-regularization and provides the mathematical framework for computing step sizes that satisfy KL-divergence bounds, addressing similar technical challenges as claimed in the original contribution.

Evidence

Evidence 1 - **Rationale**: This establishes that KL-regularized policy updates were already developed in prior work, challenging the novelty of designing such updates. - **Original**: we design a kl-regularized proximal update implementable in continuous action spaces that exponentiates the value estimate - **Candidate**: trust-region optimization for policy search was first introduced in the relative entropy policy search (reps) algorithm (peters et al. 2010). many variants of this algorithm exist (akrour et al. 2016; abdolmaleki et al. 2015;d a n i e le ta l . 2016; akrour et al. 2018). all these algorithms use a bo...

Evidence 2 - **Rationale**: This shows the candidate paper derives the exponential form of policy updates under KL-regularization for continuous spaces, demonstrating prior work on this specific technical approach. - **Original**: a kl-regularized, natural-gradient policy step exponentiates the evaluated action-value, recovering a ppo/trpo-style proximal update in continuous state-action spaces - **Candidate**: π(a|s) ∝ πold(a|s)exp (ψ(s, a)t w - eπ(·|s) [ψ(s,·)t w] η) ∝ exp (ψ(s, a)t (θold + η-1w))

---

6. A KL-regularization framework for learning to plan with adaptive priors

URL: View paper

Brief Assessment

KL Regularization Adaptive Priors[30] focuses on model-based RL with MPPI planning and uses KL regularization to align a learned sampling policy with a planner-induced prior. The original paper's contribution concerns KL-regularized proximal updates in a function-space (RKHS) framework with kernel TD critics and explicit sample complexity bounds for policy evaluation, which differs from the candidate's planning-centric approach.

Contribution 3: Instance-adaptive convergence guarantees unifying multiple RKHS regimes

Description: The authors establish non-asymptotic convergence guarantees that depend on RKHS entropy and unify multiple function-approximation regimes including tabular, linear, Sobolev spaces, Gaussian kernels, and Neural Tangent Kernels. They derive a sampling rule for the proximal update ensuring optimal k to the power negative one-half convergence rate for stochastic optimization.

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

---

1. Adaptive penalized M-estimation with current status data

URL: View paper

Brief Assessment

Adaptive Penalized M Estimation[41] focuses on current status data with penalized M-estimation in Sobolev spaces, not reinforcement learning or policy optimization in RKHS. The technical contexts are fundamentally different.

---

2. Stochastic Neural Tangent Kernel: Revisiting the NTK For SGD

URL: View paper

Brief Assessment

Stochastic Neural Tangent Kernel[37] focuses on incorporating SGD's minibatch noise into the NTK framework for stochastic optimization, not on instance-adaptive convergence guarantees unifying tabular, linear, Sobolev, Gaussian, and NTK regimes for policy optimization in RL.

---

3. Convergence and Sketching-Based Efficient Computation of Neural Tangent Kernel Weights in Physics-Based Loss

URL: View paper

Brief Assessment

Sketching Neural Tangent Kernel[40] focuses on efficient computation of NTK weights in physics-informed neural networks for multi-objective optimization, not on instance-adaptive convergence guarantees for policy optimization across multiple RKHS regimes (tabular, linear, Sobolev, Gaussian, NTK) as claimed in the original paper.

---

4. Random smoothing regularization in kernel gradient descent learning

URL: View paper

Brief Assessment

Random Smoothing Kernel Gradient[33] focuses on kernel gradient descent learning with random smoothing regularization in classical Sobolev spaces, not on policy optimization or temporal-difference learning in reinforcement learning contexts.

---

5. Identifying good directions to escape the ntk regime and efficiently learn low-degree plus sparse polynomials

URL: View paper

Brief Assessment

Escape NTK Regime[35] focuses on escaping the Neural Tangent Kernel regime through second-order Taylor expansions for learning low-degree plus sparse polynomials, not on instance-adaptive convergence guarantees across multiple RKHS regimes (tabular, linear, Sobolev, Gaussian, NTK) for policy optimization.

---

6. Primer of adaptive finite element methods

URL: View paper

Brief Assessment

Adaptive Finite Elements Primer[36] focuses on adaptive finite element methods for PDEs, not reinforcement learning or RKHS-based policy optimization. The candidate discusses multiresolution analysis, wavelets, and adaptive mesh refinement—fundamentally different from the original paper's RL convergence guarantees.

---

7. The Three Paradigms of Physics-Informed Learning: Neural Networks (PINNs), Neural Operators (PINOs), and Reinforcement Learning (PIRL)

URL: View paper

Brief Assessment

Three Paradigms Physics Informed[39] focuses on physics-informed learning paradigms (PINNs, PINOs, PIRL) with neural network error bounds beyond the NTK regime and physics residual control in dual Sobolev norms. This is fundamentally different from the original paper's focus on RKHS-based policy optimization with instance-adaptive convergence guarantees unifying tabular, linear, Sobolev, Gaussian, and NTK regimes for reinforcement learning.

---

8. Adaptive Optimization in the -Width Limit

URL: View paper

Brief Assessment

Adaptive Optimization Width Limit[34] focuses on adaptive optimizers (Adam, etc.) in neural networks' infinite-width limits, not on instance-adaptive convergence guarantees for temporal-difference learning across RKHS regimes as in the original paper.

---

9. Model-Robust and Adaptive-Optimal Transfer Learning for Tackling Concept Shifts in Nonparametric Regression

URL: View paper

Brief Assessment

Model Robust Transfer Learning[38] focuses on transfer learning for nonparametric regression with concept shifts using Gaussian kernels in Sobolev spaces, not on policy optimization or temporal-difference learning in reinforcement learning settings.

Appendix: Text Similarity Detection

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

References

- [0] Sampling Complexity of TD and PPO in RKHS View paper
- [1] Gaussian process temporal-difference learning with scalability and worst-case performance guarantees View paper
- [2] Residual kernel policy network: Enhancing stability and robustness in rkhs-based reinforcement learning View paper
- [3] Online Attentive Kernel-Based Temporal Difference Learning View paper
- [4] Statistical Efficiency of Distributional Temporal Difference Learning and Freedman's Inequality in Hilbert Spaces View paper
- [5] A non-asymptotic analysis of non-parametric temporal-difference learning View paper
- [6] Optimal policy evaluation using kernel-based temporal difference methods View paper
- [7] Kernel-based decentralized policy evaluation for reinforcement learning View paper
- [8] Kernel temporal difference based reinforcement learning for brain machine interfaces View paper
- [9] Online selective kernel-based temporal difference learning View paper
- [10] Kernel-based least squares temporal difference with gradient correction View paper
- [11] Finite-sample analysis of multi-agent policy evaluation with kernelized gradient temporal difference View paper
- [12] Kernel temporal differences for neural decoding View paper
- [13] Towards Specialized Reinforcement Learning From Diverse Data View paper
- [14] Regularized policy iteration View paper
- [15] Kernel temporal differences for reinforcement learning with applications to brain machine interfaces View paper
- [16] An efficient L2-norm regularized least-squares temporal difference learning algorithm View paper
- [17] A sparse kernel-based least-squares temporal difference algorithm for reinforcement learning View paper
- [18] Breaking bellman's curse of dimensionality: Efficient kernel gradient temporal difference View paper
- [19] Kernel Temporal Differences for EEG-based Reinforcement Learning Brain Machine Interfaces. View paper
- [20] CIM-PPO:Proximal Policy Optimization with Liu-Correntropy Induced Metric View paper
- [21] Reinforcement learning in partially observable environments using approximate information state View paper
- [22] Reinforcement learning via kernel temporal difference. View paper
- [23] Policy Evaluation in Continuous MDPs with Efficient Kernelized Gradient Temporal Difference View paper
- [24] PPO-CIM: Proximal Policy Optimization with Correntropy Induced Metric View paper
- [25] Gain Function Tracking in the Feedback Particle Filter View paper
- [26] RKHS Temporal Difference Learning View paper
- [27] Optimal Rates of Convergence for Entropy Regularization in Discounted Markov Decision Processes View paper

- [28] Achieving Zero Constraint Violation for Constrained Reinforcement Learning via Conservative Natural Policy Gradient Primal-Dual Algorithm View paper
- [29] Compatible natural gradient policy search View paper
- [30] A KL-regularization framework for learning to plan with adaptive priors View paper
- [31] FCPortugal-multi-robot action learning View paper
- [32] SACrificing Intuition View paper
- [33] Random smoothing regularization in kernel gradient descent learning View paper
- [34] Adaptive Optimization in the -Width Limit View paper
- [35] Identifying good directions to escape the ntk regime and efficiently learn low-degree plus sparse polynomials View paper
- [36] Primer of adaptive finite element methods View paper
- [37] Stochastic Neural Tangent Kernel: Revisiting the NTK For SGD View paper
- [38] Model-Robust and Adaptive-Optimal Transfer Learning for Tackling Concept Shifts in Nonparametric Regression View paper
- [39] The Three Paradigms of Physics-Informed Learning: Neural Networks (PINNs), Neural Operators (PINOs), and Reinforcement Learning (PIRL) View paper
- [40] Convergence and Sketching-Based Efficient Computation of Neural Tangent Kernel Weights in Physics-Based Loss View paper
- [41] Adaptive penalized M-estimation with current status data View paper