Designing Theoretically Principled and Practical Actor-Critic Algorithms

by

Reza Asad

M.Sc., Simon Fraser University, 2020
B.Sc., University of Toronto, 2013

Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of
Doctor of Philosophy

in the
Department of Computing Science
Faculty of Applied Sciences

© Reza Asad 2026
SIMON FRASER UNIVERSITY
Fall 2026

Copyright in this work is held by the author. Please ensure that any reproduction or re-use is done in accordance with the relevant national copyright legislation.

Declaration of Committee

Name:	**Reza Asad**
Degree:	**Doctor of Philosophy**
Thesis title:	**Designing Theoretically Principled and Practical Actor-Critic Algorithms**
Committee:	**Chair:** Hedayat Zarkoob Lecturer, Computing Science
	Sharan Vaswani Supervisor Assistant Professor, Computing Science
	Manolis Savva Supervisor Associate Professor, Computing Science
	Jason Peng Committee Member Assistant Professor, Computing Science
	Wuyang Chen Internal Examiner Assistant Professor, Computing Science
	Amir-massoud Farahmand External Examiner Associate Professor, Computing Science

Table of Contents

Declaration of Committee ii

Table of Contents iii

List of Tables vii

List of Figures viii

Abstract x

Preface xi

Acknowledgements xiii

1 Introduction 1

1.1 Policy Gradient Methods in RL 2
1.2 Actor-Critic Methods and Off-Policy Learning 3
1.3 Optimistic Actor-Critic Methods with Parameterized Actors 4
1.4 Thesis Contributions 6

2 Fast Convergence of Softmax Policy Mirror Ascent 9

2.1 Introduction 9
2.2 Problem Formulation 11
2.3 Softmax Policy Mirror Ascent: Tabular Parametrization 11
2.3.1 Bandit Setting 12
2.3.2 MDP Setting 12
2.3.3 Theoretical Results 13
2.4 Handling Function Approximation 15
2.4.1 Comparison to Existing Approaches 16
2.4.2 Theoretical Guarantee 17
2.5 Empirical Evaluation 19
2.6 Discussion 22

3 Dissecting Discrete Soft Actor-Critic: Limitations and Principled Alternatives **23**
3.1 Introduction . . . 23
3.2 Preliminaries . . . 26
3.3 General Off-policy Actor-critic Framework . . . 26
3.3.1 Policy Evaluation . . . 26
3.3.2 Policy Optimization . . . 27
3.4 Theoretical Guarantees . . . 30
3.5 Empirical Evaluation . . . 32
3.6 Conclusion and Future Work . . . 36

4 Optimistic Actor-Critic with Parametric Policies **37**
4.1 Introduction . . . 37
4.2 Preliminaries . . . 40
4.3 A General Actor-Critic Framework with Parametric Policies . . . 41
4.4 Instantiating the Actor: Projected Natural Policy Gradient . . . 42
4.4.1 Projected Natural Policy Gradient . . . 42
4.4.2 Controlling the Projection Error for Log-Linear Policies . . . 44
4.4.3 Putting Everything Together: Projected NPG with Log-Linear Policies 46
4.5 Instantiating the Critic: Langevin Monte Carlo . . . 47
4.5.1 LMC for Linear MDPs . . . 47
4.5.2 Optimism Guarantee and Error Bound . . . 48
4.6 Sample Complexity Analysis . . . 50
4.6.1 On-Policy Setting . . . 50
4.6.2 Off-Policy Setting . . . 51
4.7 Experiments . . . 52
4.7.1 Experiments in Linear MDPs . . . 53
4.7.2 Experiments Beyond Linear MDPs: Atari . . . 53
4.8 Discussion . . . 55

5 Conclusion **56**
5.1 Summary of Contributions . . . 56
5.2 Future Work . . . 57

Bibliography **59**

Appendix A Fast Convergence of Softmax Policy Mirror Ascent **70**
A.1 Multi-armed Bandit Proofs . . . 70
A.1.1 Super-linear Rate for Bandits . . . 72
A.2 MDP Proofs . . . 74

A.2.1 Tabular Setting . . . 74
A.2.2 Function Approximation With Exact Advantage . . . 78
A.2.3 Function Approximation With Inexact Advantage . . . 80
A.3 Tabular MDP Experiments . . . 86
A.3.1 Softmax Tabular Representation . . . 86
A.3.2 Linear Functional Approximation (Linear FA) . . . 88
A.3.3 Empirical Verification that C_t is Non-zero . . . 89
A.4 Additional Details for Stable Baselines Experiments . . . 90
A.4.1 Atari and Mujoco Details . . . 90
A.4.2 Ablation Study on the Number of Inner Loop Optimization Steps . . . 92

Appendix B Dissecting Discrete Soft Actor-Critic: Limitations and Principled Alternatives 96
B.1 Actor Objective Instantiations . . . 96
B.1.1 NPG-FKL(τ, ζ) . . . 96
B.1.2 SPMA-FKL(τ, ζ) . . . 97
B.1.3 NPG-RKL(τ, ζ) . . . 97
B.1.4 SPMA-RKL(τ, ζ) . . . 98
B.2 General Off-Policy Actor-Critic Pseudocode . . . 98
B.3 Theoretical Results . . . 99
B.3.1 Proof of Theorem 4 . . . 99
B.3.2 Proof of Corollary 1 . . . 101
B.3.3 Proof of Corollary 2 . . . 112
B.3.4 Proof of Theorem 5 . . . 120
B.4 Helper Lemmas . . . 125
B.5 Experimental Details and Additional Results . . . 128
B.5.1 Details for Stable Baselines Experiments . . . 128
B.5.2 Comparison to GreedyAC . . . 129
B.5.3 Does Double Q Learning Hurt or Improve Performance? . . . 130
B.5.4 Is Fixed Target Entropy the Main Bottleneck in DSAC? . . . 132
B.5.5 Evaluating our DSAC Variant on Hard Exploration Games . . . 133
B.5.6 Evaluating our DSAC Variant on stable-retro games . . . 134
B.5.7 Our Objectives Benefit Similarly from the Hard Bellman Operator . . . 135
B.5.8 Additional Results: Is Entropy Regularization Necessary? . . . 137
B.5.9 Expected reward after 10M steps . . . 140
B.5.10 Full Horizon Training Curves Across 20 Games . . . 141
B.5.11 Entropy Collapse of Default DSAC . . . 144

Appendix C Optimistic Actor-Critic with Parametric Policies 145
C.1 Notations . . . 145

C.2 Analysis for the Actor . . . 146
C.2.1 Generalized OMD Regret (Proof of Theorem 6) . . . 146
C.2.2 Projection Error (Proof of Lemma 1) . . . 147
C.2.3 Instantiating the Actor with SPMA . . . 150
C.2.4 Technical Tools . . . 152
C.3 Constructing $\mathcal{D}_{\text{exp}}$ and ρ_{exp} via Experimental Design . . . 153
C.3.1 Kiefer–Wolfowitz Theorem and G-Experimental Design . . . 153
C.3.2 Exploratory Policy and Minimum Eigenvalue . . . 153
C.4 Analysis for the Critic . . . 155
C.4.1 Proof of Lemma 2 . . . 155
C.4.2 Proofs of Preliminary Properties . . . 159
C.4.3 Technical Tools . . . 173
C.5 Sample Complexity in the On-Policy Setting . . . 174
C.5.1 Proof of Good Event . . . 174
C.5.2 Proof of Theorem 7 . . . 175
C.5.3 Technical Tools . . . 178
C.6 Sample Complexity in the Off-Policy Setting . . . 179
C.6.1 Covering Number (Proof of Lemma 3) . . . 179
C.6.2 Proof of Good Event . . . 181
C.6.3 Proof of Theorem 8 . . . 182
C.6.4 Technical Tools . . . 184
C.7 Experiments . . . 184
C.7.1 Experiments in Linear MDPs . . . 184
C.7.2 Ablation Studies . . . 187
C.7.3 Atari Experimental Details . . . 188

List of Tables

Table 3.1 Actor objectives in the proposed off-policy framework. 29
Table 3.2 Zero-entropy IQM scores for proposed off-policy objectives. 34
Table 3.3 IQM scores with actor entropy and no critic entropy. 35

Table 4.1 Comparison with prior linear MDP actor-critic algorithms. 38

Table A.1 Hyperparameters for Atari experiments. 90
Table A.2 Hyperparameters for MuJoCo experiments. 91

Table B.1 Hyperparameters for Atari and Retro experiments. 128
Table B.2 IQM comparison of DSAC with zero critic entropy and SD-SAC. . . . 130
Table B.3 IQM comparison of fixed-entropy DSAC and TES-SAC. 132
Table B.4 Hard-exploration IQM scores for DSAC with zero critic entropy and baselines. 133
Table B.5 IQM comparison of zero and constant critic entropy in DSAC. 135

Table C.1 Notation used for the actor-critic framework. 145
Table C.2 Hyperparameter search space for linear MDP experiments. 186
Table C.3 On-policy Atari hyperparameters. 190
Table C.4 Off-policy Atari hyperparameters. 190

List of Figures

Figure 1.1 N-chain illustration of the exploration challenge. 5

Figure 2.1 SPMA matches or outperforms baselines on Atari games. 20
Figure 2.2 SPMA performs competitively on MuJoCo control tasks. 21

Figure 3.1 Removing critic entropy improves DSAC on Atari games. 24
Figure 3.2 DSAC performance is sensitive to critic entropy. 33

Figure 4.1 LMC-NPG-EXP performance in the Random MDP. 53
Figure 4.2 On-policy Atari comparison between LMC-NPG-EXP and PPO. . . 54
Figure 4.3 Off-policy Atari comparison between LMC-NPG-EXP and DQN. . 54

Figure A.1 SPMA matches NPG and outperforms SPG. 86
Figure A.2 SPG is more sensitive to the stepsize than SPMA and NPG. 87
Figure A.3 CliffWorld convergence under varying optimization error and bias. . 88
Figure A.4 FrozenLake convergence under varying optimization error and bias. 88
Figure A.5 Softmax tabular SPMA yields a positive convergence constant. . . . 89
Figure A.6 MuJoCo results for larger inner-loop optimization budgets. 92
Figure A.7 Atari results for larger inner-loop optimization budgets. 93
Figure A.8 MuJoCo ablation on the number of inner-loop optimization steps. . 94
Figure A.9 Atari ablation on the number of inner-loop optimization steps. . . . 95

Figure B.1 GreedyAC comparison with DSAC using batch size 32. 129
Figure B.2 GreedyAC comparison with DSAC using batch size 256. 129
Figure B.3 Per-game Atari rewards for DSAC with zero critic entropy and SD-SAC. 131
Figure B.4 Per-game Atari rewards for DSAC with zero critic entropy and TES-SAC. 132
Figure B.5 Per-game hard-exploration rewards for DSAC with zero critic entropy and baselines. 133
Figure B.6 Stable-Retro performance of DSAC without critic entropy. 134
Figure B.7 Effect of disabling critic entropy for KL-based objectives. 135
Figure B.8 Best constant critic-entropy values across Atari games. 136
Figure B.9 Policy entropy improves with more actor optimization steps. 137

Figure B.10 More actor optimization steps improve zero-entropy objectives. . . 137
Figure B.11 KL divergence stabilizes with more actor optimization steps. 138
Figure B.12 Objectives from the proposed framework remain robust without entropy. 139
Figure B.13 Per-game Atari rewards for proposed objectives and baselines. . . . 140
Figure B.14 Full-horizon comparison between PPO and proposed objectives. . . . 141
Figure B.15 Full-horizon comparison between DQN and proposed objectives. . . 142
Figure B.16 Full-horizon comparison between LMC and proposed objectives. . . 143
Figure B.17 Policy entropy collapse explains poor default DSAC performance. . 144

Figure C.1 Deep Sea environment example. 185
Figure C.2 LMC-NPG-EXP performance in the Deep Sea environment. 187
Figure C.3 Ablation of the exploration mechanism in LMC-NPG-EXP. 187
Figure C.4 Effect of feature dimension in Deep Sea. 188
Figure C.5 LMC-NPG-EXP is robust to inverse-temperature choices. 188
Figure C.6 LMC-NPG-EXP is robust to choices of M. 188

Abstract

Reinforcement learning studies how agents learn to make sequential decisions by interacting with an environment, receiving feedback, and using that feedback to improve their behavior over time. A central challenge is that an agent must learn effectively from the experience it has already collected, while also choosing actions that generate useful data for subsequent learning. This thesis studies policy-gradient and actor-critic methods that address different aspects of this challenge while aiming to be both theoretically principled and practical. First, it introduces Softmax Policy Mirror Ascent (`SPMA`), a logit-space mirror-ascent policy-gradient method with convergence guarantees in tabular and function-approximation settings. The method achieves strong empirical performance in on-policy experiments on tabular Markov decision processes, Atari, and MuJoCo. Second, it turns to discrete-action off-policy actor-critic methods. In this setting, it identifies critic entropy regularization as a key bottleneck in Discrete Soft Actor-Critic and develops a modular framework for designing more robust actor-critic objectives. The framework includes an off-policy extension of `SPMA`, which is empirically evaluated and shown to perform well in this setting. The thesis also establishes tabular convergence guarantees for the broader class of objectives induced by the framework. Finally, the thesis directly addresses exploration, which is not explicitly handled by the preceding policy-optimization and actor-critic frameworks. It studies optimistic actor-critic methods for finite-horizon linear Markov decision processes with explicit parametric policies. The approach combines logit matching with Langevin-based optimistic value estimation, leading to end-to-end sample-complexity guarantees. The method is evaluated on linear MDP experiments that match the theoretical setting and illustrate the benefit of optimism with parametrized actors. The thesis also presents preliminary practical adaptations of this optimistic actor-critic algorithm for large-scale Atari experiments.

Keywords: Reinforcement Learning; Policy Gradient; Mirror Ascent; Off-Policy Discrete-Action Actor-Critic; Entropy Regularized Policy Optimization; Optimism; Linear MDPs

Preface

The main technical material in this thesis is drawn from two published papers and one manuscript currently under review. Chapters 2, 3, and 4 reproduce the corresponding papers with only minor changes to formatting for consistency within the thesis. The remaining material, including the abstract, introductory material, and concluding discussion, was written specifically for this thesis.

- Chapter 2 appeared in the Proceedings of the International Conference on Artificial Intelligence and Statistics, 2025 [Asad et al., 2025b]:

  Asad, Reza, Reza Babanezhad Harikandeh, Issam H. Laradji, Nicolas Le Roux, and Sharan Vaswani. "Fast Convergence of Softmax Policy Mirror Ascent." In Proceedings of the International Conference on Artificial Intelligence and Statistics, 2025.

  Reza Asad was the major contributor in terms of writing, algorithmic design, and empirical results. Sharan Vaswani and Reza Babanezhad were the major contributors to the theoretical proofs. Issam Laradji provided valuable feedback on designing an efficient pipeline for running experiments. Reza Babanezhad, Nicolas Le Roux, and Sharan Vaswani provided valuable feedback on the paper.

- Chapter 3 is under review and is available as a preprint [Asad et al., 2025a]:

  Asad, Reza, Reza Babanezhad, and Sharan Vaswani. "Dissecting Discrete Soft Actor-Critic: Limitations and Principled Alternatives." arXiv preprint arXiv:2509.09838, 2025.

  Reza Asad was the major contributor in terms of writing, the design of the theoretical objectives, and empirical results. Sharan Vaswani and Reza Babanezhad were the major contributors to the theoretical proofs and provided valuable feedback on the paper.

- Chapter 4 appeared in the Proceedings of the International Conference on Artificial Intelligence and Statistics, 2026 [Lin et al., 2026]:

  Lin, Max Qiushi, Reza Asad, Kevin Tan, Haque Ishfaq, Csaba Szepesvári, and Sharan Vaswani. "Optimistic Actor-Critic with Parametric Policies for Linear Markov Decision

Processes." In Proceedings of the International Conference on Artificial Intelligence and Statistics, 2026.

Qiushi Lin was the major contributor in terms of writing and theoretical results. Reza Asad was the major contributor to the empirical results. Sharan Vaswani and Kevin Tan contributed to the theoretical results. Kevin Tan, Haque Ishfaq, Csaba Szepesvári, and Sharan Vaswani provided valuable feedback on the paper.

Acknowledgements

I am deeply grateful for the guidance, support, and kindness of many people throughout my PhD journey. I am especially thankful to my advisors, Sharan Vaswani and Manolis Savva. I began my PhD working with Manolis in 3D vision, and only later realized that my interests were more strongly drawn toward optimization for machine learning and reinforcement learning. I am profoundly grateful to Manolis for supporting me throughout this transition and for continuing to guide me with generosity and understanding. His kindness had a lasting impact on me and shaped the way I think about mentorship.

I am equally grateful to Sharan Vaswani. I met Sharan when I took his course on Optimization in Machine Learning, and without this course, I may not have found the research direction that I am deeply interested in. I feel incredibly lucky to have learned from his deep knowledge, exceptional ability to explain difficult concepts, and sharp instinct for important research problems, all of which shaped my research mindset. I am sincerely thankful for his mentorship, his continuous support, and the opportunities he helped me find to pursue internships and collaborate with brilliant researchers. His encouragement and thoughtful questions always pushed me to think more deeply and to engage with ideas I might not have approached on my own. I look forward to continuing to collaborate with him in the future.

I would also like to thank my collaborators. I am grateful to Reza Babanezhad for his valuable feedback, sharp theoretical understanding, and ability to bring new perspectives to our projects. I am deeply thankful to Nicolas Le Roux for his incredible mentorship during my internship at Microsoft. I feel very fortunate to have worked with someone so genuine, kind, and exceptionally insightful. His guidance helped me push myself intellectually, and his continued support during my search for future opportunities meant a great deal to me. I would also like to thank Max Qiushi Lin for his outstanding effort in pushing the final project of this thesis forward and helping bring it to completion. I am grateful to Issam Laradji for his valuable guidance in helping me develop a better mindset for running experiments efficiently, and also for organizing many beach BBQs that gave me much-needed breaks from research when I was stuck.

I am also thankful to my friends for their support, patience, and understanding throughout the PhD journey: Michael, Ashwin, Mohammad, Navid, Parsa, Hossein, Parnian, Saurabh,

and Chirag. Their friendship made the difficult parts of this journey easier and the good moments more meaningful.

I am also deeply grateful to my girlfriend, Malina, who came into my life during the final stages of my PhD. Her kindness, encouragement, and support, especially as I prepared for my defence, meant a great deal to me.

Finally, I would like to thank my family for their unconditional love and support. I have been incredibly lucky to have parents who have always stood by me. Our Sunday dinners gave me a place to step away from research, reset, and feel grounded. I am also deeply grateful to my sister, whose visits to Vancouver brought so much joy and energy to our family, and whose support during our weekly calls helped me throughout this journey.

Chapter 1

Introduction

In standard supervised learning, the learner fits a model to a fixed dataset of observed examples, and the process that generated those examples is usually treated as outside the algorithmic problem. This perspective is well-suited to prediction tasks, but many applications require more than predicting from fixed data. In robotics [Kober et al., 2013], control [Todorov et al., 2012], games [Bellemare et al., 2013, Mnih et al., 2015], and learning from human feedback in large language models [Ouyang et al., 2022, Uc-Cetina et al., 2023], a learned system must often make decisions over time, with consequences that extend beyond a single prediction. Reinforcement learning (RL) studies this sequential decision-making setting: agents learn through interaction with an environment and use feedback from that interaction to improve their behavior. Crucially, unlike in standard supervised learning, the agent's decisions affect both its immediate performance and the experience available for future learning.

RL problems are commonly modeled using Markov decision processes (MDPs). At each time step, the agent observes a state, which summarizes the information available for decision making. The agent then selects an action and receives feedback in the form of a reward. The reward function specifies this numerical feedback and is problem-dependent. After the action is taken, the environment moves to a new state. The transition dynamics describe how the next state is generated from the current state and action. For example, in a robotics task, the state may include sensor readings and the position of the robot, the action may specify a movement command, and the reward may measure progress toward a target. A state-action pair refers to a particular state together with a particular action available at that state.

The agent's behavior is described by a policy, which specifies how actions are selected from states. The goal is to learn a policy that maximizes expected cumulative reward, often called expected return: the average total reward accumulated over a sequence of interactions. If the transition dynamics and reward function are known, the problem is to plan a good sequence of actions. In most learning problems, however, this information is unknown and must be learned through interaction with the environment.

Because the agent's policy influences the experience it collects, policy improvement and data collection are coupled. An agent may repeatedly choose actions that appear effective based on its current experience, while failing to try alternatives that could reveal better long-term behavior. An RL algorithm must therefore use the data already collected to improve the policy, while also collecting experience that supports reliable future improvement. This coupling is one reason RL algorithms are difficult to analyze and difficult to make reliable in practice.

The rest of this introduction uses the coupling between policy optimization and data collection as a common thread for the three technical directions studied in the thesis. Section 1.1 discusses policy-gradient methods and the gap between theory and practice in policy optimization. Section 1.2 turns to actor-critic methods that can use data collected by previous policies, while combining policy optimization with estimates of the future consequences of different decisions. Section 1.3 discusses exploration and the role of optimism in actor-critic methods for collecting informative data. Section 1.4 then summarizes the thesis contributions and explains how they connect to the three technical chapters.

1.1 Policy Gradient Methods in RL

Policy gradient (PG) methods are among the most widely used approaches for policy optimization [Williams, 1992, Sutton et al., 1999, Konda and Tsitsiklis, 2000, Kakade, 2001]. These methods parameterize the policy directly and update its parameters using gradient estimates of the expected return. This optimization viewpoint makes PG methods flexible: they can represent stochastic policies and naturally incorporate function approximation. In this setting, function approximation means representing the policy with a shared parametric model, such as a neural network, rather than with a tabular collection of separate parameters for each state-action pair. Such shared representations are essential for scaling RL algorithms beyond small tabular problems to large state-action spaces. As a result, PG ideas underlie many practical deep RL algorithms, including Trust Region Policy Optimization (`TRPO`), Proximal Policy Optimization (`PPO`), and Mirror Descent Policy Optimization (`MDPO`) [Schulman et al., 2015a, 2017b, Tomar et al., 2020].

However, PG methods expose a persistent gap between theory and practice. On the theoretical side, the policy optimization objective is generally non-concave, the gradients are estimated from data generated by the current policy, and convergence guarantees often require simplified settings such as tabular parameterizations, exact gradients, or strong coverage assumptions [Agarwal et al., 2021a, Mei et al., 2020, Yuan et al., 2023, Lu et al., 2024]. On the practical side, algorithms that perform well in large-scale experiments often use surrogate objectives, trust regions, clipping, entropy regularization, and function approximation [Schulman et al., 2015a, 2017b]. These design choices are important empirically, but they also make the resulting methods hard to analyze. Consequently, theoretically elegant

PG algorithms are not commonly used in practice, while many practical PG algorithms lack sharp guarantees in realistic function approximation settings [Tomar et al., 2020].

This gap motivates the first part of the thesis. Chapter 2 asks whether one can design a PG algorithm that has convergence guarantees beyond the tabular setting, while also achieving strong performance on standard deep RL benchmarks without requiring algorithmic modifications. The resulting method, Softmax Policy Mirror Ascent (`SPMA`), is first developed in tabular settings and then extended to function approximation through a projected mirror-ascent viewpoint.

1.2 Actor-Critic Methods and Off-Policy Learning

A practical limitation of policy-gradient methods is sample efficiency. These methods are based on the gradient of the expected return with respect to the policy parameters. Since this gradient is generally not available exactly, it is typically estimated from sampled trajectories. The rewards observed along these trajectories provide a noisy signal about how the policy parameters should change to improve return. These trajectory-based policy-gradient estimates can have high variance, leading to slow learning and poor sample efficiency [Greensmith et al., 2004, Xu et al., 2017].

Actor-critic methods aim to mitigate this issue by explicitly combining policy updates with value estimation [Konda and Tsitsiklis, 1999, Peters et al., 2005, Bhatnagar et al., 2009]. In these methods, the actor represents the policy being optimized, while the critic estimates value functions that predict expected future return from states or state-action pairs; the latter are often called action-value functions. The critic is typically learned through bootstrapping, by updating it toward targets that combine observed rewards with learned estimates of future return. The actor then uses the critic's estimates to guide the policy update. Thus, actor-critic methods remain closely related to policy-gradient methods, but their behavior depends not only on the policy update itself, but also on the value estimates used to construct it. This distinction is important throughout the thesis: Chapter 2 mainly studies the policy-optimization update while treating the required value estimates as given or externally approximated, whereas Chapters 3 and 4 make critic learning explicit and study its interaction with the actor update.

This actor-critic structure appears in both on-policy and off-policy algorithms. On-policy methods update the actor and critic using data generated by the current policy, which often leads to stable learning but can be sample inefficient [Haarnoja et al., 2018b]. Off-policy methods instead reuse data generated by previous policies or exploratory behavior. This can improve sample efficiency, but it introduces a distinction between the behavior policy that generates data and the target policy being optimized [Sutton and Barto, 2018]. When off-policy learning is combined with bootstrapping and function approximation, this distinction is part of a well-known source of instability in RL [Van Hasselt et al., 2018]. A successful

example in continuous control is Soft Actor-Critic (SAC), an off-policy actor-critic algorithm based on the maximum-entropy RL framework, where the objective augments expected return with an entropy term [Haarnoja et al., 2018a].

The off-policy setting therefore raises a question complementary to the one studied in Chapter 2. Even if one has principled ways to improve a policy, how should those ideas be combined with critic learning and data reuse? This issue becomes especially delicate in discrete-action domains. Although SAC is highly effective in continuous control, its direct discrete-action counterpart, DSAC, performs poorly on standard benchmarks such as Atari [Christodoulou, 2019]. Prior work proposes heuristics that can stabilize training of DSAC [Xu et al., 2021, Zhou et al., 2022], but these modifications introduce interacting components and additional hyperparameters, making it difficult to isolate the core bottleneck. This added complexity also leaves the resulting methods difficult to analyze, and existing variants lack theoretical guarantees even in the tabular setting.

This motivates the second part of the thesis. Chapter 3 studies off-policy actor-critic methods for discrete actions by first dissecting DSAC to identify the bottleneck limiting its performance. The extensive ablation study in this chapter shows that decoupling the entropy terms in the actor and critic updates is key to improving DSAC's performance, and motivates the design of a general actor-critic framework that subsumes DSAC as a special case.

1.3 Optimistic Actor-Critic Methods with Parameterized Actors

The first two parts of the thesis focus on principled and practical approaches to policy optimization and actor-critic design. However, understanding RL also requires addressing exploration in a systematic manner. Many analyses of PG and actor-critic methods make assumptions that ensure sufficient coverage of the relevant state-action space. For instance, convergence rates for PG methods often depend on bounded distribution-mismatch coefficients [Agarwal et al., 2021a, Yuan et al., 2023, Alfano and Rebeschini, 2022b, Asad et al., 2025b]. Similarly, analyses of actor-critic methods may assume reachability of the state-action space, coverage of the collected data, bounded mismatch ratios, or mixing of the induced Markov chain [Abbasi-Yadkori et al., 2019, Neu et al., 2017, Bhandari and Russo, 2024, Agarwal et al., 2021a, Cen et al., 2022, Gaur et al., 2024, Hong et al., 2023, Fu et al., 2021, Xu et al., 2020, Cayci et al., 2024, Khodadadian et al., 2022, Gaur et al., 2023]. These assumptions are useful for isolating the policy-optimization problem, but they do not explain how the algorithm obtains informative data in the first place.

Alternatively, entropy regularization can be used to encourage exploration by promoting action diversity, as in the actor-critic framework studied in Chapter 3. However, the theoretical analysis in that chapter assumes one can obtain a good estimate of the action-value function for every state-action pair in the tabular setting, and the exploration problem is assumed

away. Moreover, in some environments (e.g., the N-chain), entropy regularization alone need not ensure efficient exploration [Cesa-Bianchi et al., 2017, O'Donoghue, 2023].

The illustrative N-chain in Figure 1.1 makes this distinction concrete. Starting at s_1, the agent can either advance to the next state, incurring a small cost, or reset to s_1 with zero reward. A positive reward is available only upon reaching s_N, and the cost of advancing is small enough that reaching this terminal state is preferable to repeatedly resetting. Suppose the agent has explored an initial prefix of the chain but has little information about the remaining states. Encouraging action diversity within the visited prefix does not by itself favour the sustained sequence of advances needed to discover the distant reward. Before that reward has been discovered, the immediate costs of advancing may make resetting appear preferable. The agent may therefore repeatedly revisit familiar states while the value of the continuation remains uncertain.

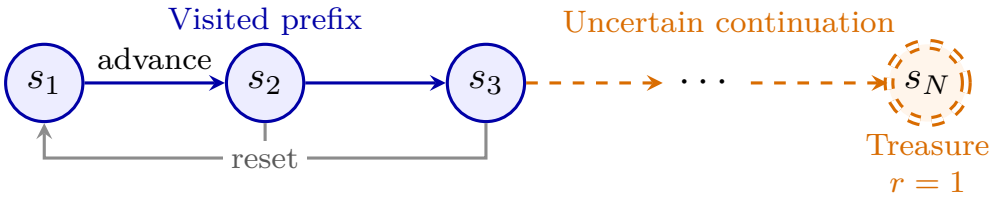


Figure 1.1: An illustrative N-chain. Advancing incurs a small cost, resetting to s_1 yields zero reward, and reaching s_N yields a positive terminal reward. The solid portion represents the visited prefix; the dashed portion represents the uncertain continuation. The terminal reward is shown for illustration and is unknown to the learner until discovered.

Optimism addresses this issue by using uncertainty in estimated values to favour actions with high plausible returns. In the chain, an optimistic estimate of the value of the uncertain continuation can make advancing attractive despite its immediate cost, even before the terminal reward has been observed. Propagating these estimates back through the visited states allows the critic to guide the actor toward further exploration. This motivates studying how to collect informative experience while learning the actor and critic, with guarantees that account for exploration explicitly.

Finite-horizon linear MDPs provide a tractable setting for studying exploration. In this setting, the agent is given a feature representation for each state-action pair, and the unknown rewards and transition dynamics are assumed to depend linearly on these features. This structure allows data collected at one state-action pair to inform value estimates at other state-action pairs with related features. At the same time, exploration remains nontrivial: the learner must collect trajectories that are informative for the features that affect long-term return. Recent work develops optimistic algorithms for this exploration problem in linear MDPs [Cai et al., 2020, Jin et al., 2021, Zanette et al., 2021, Zhong and Zhang, 2023, Agarwal et al., 2023, He et al., 2023, Liu et al., 2023, Sherman et al., 2023, Cassel and Rosenberg, 2024, Tan et al., 2025]. However, many of these algorithms differ

from practical actor-critic methods. Some use natural policy-gradient (NPG) updates that define implicit policies from stored action-value functions, while others rely on confidence bonuses or specialized algorithmic components that complicate implementation. In contrast, practical actor-critic methods typically use explicit parametric policies and gradient-based updates.

This creates another theory-practice gap. Chapter 4 addresses this gap by designing an optimistic actor-critic algorithm for finite-horizon linear MDPs with an explicitly parameterized actor. The chapter empirically evaluates the proposed algorithm on linear MDPs that match the theoretical setting. It also considers practical on-policy and off-policy Atari implementations as a preliminary step toward adapting these ideas beyond the linear MDP setting.

Together, these challenges motivate the central theme of this thesis: designing policy-gradient and actor-critic algorithms that preserve practical algorithmic structure while admitting clearer theoretical guarantees. The next section summarizes the three contributions of the thesis.

1.4 Thesis Contributions

This thesis makes three contributions, corresponding to Chapters 2, 3, and 4.

Chapter 2: A practical on-policy PG method with convergence guarantees. The first contribution is Softmax Policy Mirror Ascent (SPMA), an on-policy PG algorithm motivated by the gap between theoretically analyzed PG methods and practical algorithms such as TRPO, PPO, and MDPO. The key idea is to perform mirror ascent over the logits using the log-sum-exp mirror map, whose gradient maps logits to action probabilities. This yields an update that can be expressed directly in probability space. This contrasts with NPG, which can be viewed as mirror ascent over the probability simplex and leads to an exponentiated-gradient update requiring explicit normalization across actions. Under appropriate conditions, the SPMA update avoids this explicit normalization. In the tabular setting, SPMA with a constant step-size achieves linear convergence under exact policy gradients, matching the rate of NPG and improving over the sublinear behavior of (accelerated) softmax policy gradient with constant step-sizes.

The chapter then extends SPMA to the function approximation setting by interpreting the policy class as a constraint on the logits. This yields a projected mirror-ascent method for log-linear and neural policy classes. For log-linear policies, the SPMA update reduces to solving convex softmax classification problems, which enables the analysis to prove convergence to a neighborhood of the optimal value function at a linear rate. This optimization structure contrasts with MDPO, a practical function-approximation extension of NPG, whose surrogate can be non-convex even under tabular softmax parameterization. Empirically, SPMA is evaluated

on simple MDPs with tabular and linear policy parameterizations, MuJoCo tasks with MLP policies, and Atari games with convolutional policies. Across these benchmarks, it matches or outperforms strong baselines such as PPO, TRPO, and MDPO.

Chapter 3: A principled off-policy actor-critic framework for discrete actions. This chapter studies how policy-optimization ideas interact with critic learning and data reuse in the off-policy discrete-action setting. It is motivated by the gap between the empirical success of SAC in continuous-control domains and the comparatively weak performance of its discrete-action counterpart, discrete SAC (DSAC), on Atari. Through a detailed empirical dissection, the chapter identifies critic entropy as a key bottleneck: removing entropy regularization from the critic update, while retaining the entropy-regularized actor update and automatic entropy tuning, substantially improves performance without requiring algorithmic changes.

Motivated by this finding, the chapter introduces a modular off-policy actor-critic framework for discrete actions. The framework decouples actor entropy from critic entropy, allowing asymmetric designs not captured by prior frameworks. This view subsumes DSAC and yields novel objectives by splitting the actor update into two steps. The first is to compute an intermediate policy using NPG or SPMA. The second is to project this policy back to the realizable policy class using either forward or reverse KL. In particular, the objective based on the SPMA intermediate policy and forward KL projection can be viewed as the off-policy extension of the algorithm developed in Chapter 2.

The chapter also provides tabular convergence guarantees for the proposed actor-critic framework, giving the first analysis that accommodates asymmetric actor and critic entropy. The analysis reduces convergence to two components: the policy evaluation error of the critic and the regret associated with the actor update. Empirically, the proposed objectives are evaluated on Atari. The results show that, like DSAC, these objectives benefit from hard Bellman policy evaluation, while unlike DSAC, they can remain robust even without entropy regularization or explicit exploration.

Chapter 4: An optimistic actor-critic algorithm with parametric policies. This chapter addresses exploration, which is often handled in PG and actor-critic analyses through coverage or bounded-mismatch assumptions. Instead of assuming that the relevant state-action space is already sufficiently covered, the chapter studies exploration in finite-horizon linear MDPs and develops an optimistic actor-critic algorithm with an explicit policy parameterization.

The actor uses a log-linear policy class and is trained through a logit-matching regression objective. This objective approximates the implicit policy induced by NPG while maintaining an explicit parametric policy whose action probabilities can be computed efficiently. The critic uses Langevin Monte Carlo to produce optimistic value estimates through a gradient-based

update with injected Gaussian noise, avoiding explicit upper-confidence-bound bonuses. This design keeps the algorithm closer to practical actor-critic implementations while retaining the optimism needed for efficient exploration.

The resulting algorithm has end-to-end sample-complexity guarantees without bounded mismatch or data-coverage assumptions. In the on-policy setting, it learns an ϵ-optimal policy with $\widetilde{\mathcal{O}}(\epsilon^{-4})$ samples. In the off-policy setting, it achieves $\widetilde{\mathcal{O}}(\epsilon^{-2})$ sample complexity. These rates match prior state-of-the-art theoretical results while using an explicitly parameterized actor and a comparatively practical critic update.

Beyond the theoretical analysis, this chapter empirically evaluates the proposed algorithm on linear MDPs that match the setting of the theory. These experiments examine whether the optimistic actor-critic design behaves as expected in the regime covered by the analysis, and illustrate the benefit of incorporating exploration when using explicitly parameterized actors. As a preliminary step toward adapting these ideas beyond linear MDPs, the chapter also presents practical on-policy and off-policy Atari variants of the method.

Chapter 2

Fast Convergence of Softmax Policy Mirror Ascent

2.1 Introduction

Policy gradient (PG) methods [Williams, 1992, Sutton et al., 1999, Konda and Tsitsiklis, 2000, Kakade, 2001] have been critical to the achievements of reinforcement learning (RL). Although the PG objective is non-concave, recent theoretical research [Agarwal et al., 2021a, Mei et al., 2020, 2021b, Bhandari and Russo, 2021, Lan, 2023, Shani et al., 2020, Liu et al., 2024, Lu et al., 2024, Alfano and Rebeschini, 2022a, Yuan et al., 2023] has analyzed PG methods in simplified settings and demonstrated their global convergence to an optimal policy. While such simplified analyses are helpful in understanding the underlying optimization issues, the resulting methods are rarely used in practice. On the other hand, while methods such as `TRPO` [Schulman et al., 2015a], `PPO` [Schulman et al., 2017b], `MDPO` [Tomar et al., 2020] are commonly used in deep RL, their theoretical analysis in the function approximation setting is quite limited. In particular, existing work either (i) analyzes these methods only in the impractical tabular setting [Tomar et al., 2020, Shani et al., 2020] or (ii) modifies these algorithms to make them more amenable to theoretical analysis [Liu et al., 1906, Zhong and Zhang, 2024]. Unfortunately, these modified algorithms are quite different from the original variants and are not systematically benchmarked on standard environments. Consequently, there exists a large gap between PG methods that have theoretical guarantees in realistic settings versus those which are implemented in practice. To make matters worse, it has been demonstrated that code-level implementation details impact the empirical performance more than the underlying algorithm [Engstrom et al., 2019].

Designing theoretically principled PG algorithms that simultaneously have good empirical performance on the standard set of benchmarks is the main motivation behind this work. To that end, we leverage an algorithm first proposed by Vaswani et al. [2021], which we modify to remove the need for normalization. We coin this refinement **S**oftmax **P**olicy **M**irror **A**scent (referred to as `SPMA`). We show that `SPMA` has comparable convergence guarantees as

existing theoretical techniques [Lu et al., 2024, Yuan et al., 2023] in the tabular and function approximation settings, while achieving comparable practical performance as PPO, TRPO and MDPO, without additional algorithmic modifications. In particular, we make the following contributions.

Contribution 1: In Section 2.3, we focus on the multi-armed bandit and tabular MDP settings, where the number of parameters scales with the number of states and actions. We develop the SPMA algorithm, which parameterizes the policy using the softmax function and uses a mirror ascent (with the log-sum-exp mirror map) update. Compared to NPG that can be viewed as mirror ascent in the space of probabilities, SPMA corresponds to mirror ascent in the dual space of logits and does not require a normalization across actions. Given access to the exact policy gradients, we prove that SPMA with a constant step-size converges to the optimal policy at a linear rate and thus matches the rate of NPG [Khodadadian et al., 2021, Liu et al., 2024]. In comparison, constant step-size softmax policy gradient (SPG) [Agarwal et al., 2021a, Mei et al., 2020] can only achieve sublinear convergence rates even with Nesterov acceleration [Chen et al., 2023]. Hence, by changing the mirror map (from Euclidean to log-sum-exp) while using the same policy parameterization, SPMA can result in an exponential improvement over SPG.

Contribution 2: In order to handle MDPs with large state-action spaces, we use function approximation (e.g. linear models or neural networks) to parameterize the policies resulting in the class of log-linear or energy-based policies [Haarnoja et al., 2017, Agarwal et al., 2021a, Yuan et al., 2023] respectively. By interpreting the policy parameterization as a constraint on the corresponding logits, we use projected mirror ascent to extend SPMA to the FA setting and design Algorithm 1. Unlike that for NPG, generalizing SPMA does not require compatible function approximation, and thus results in a more practical algorithm. Unlike MDPO [Tomar et al., 2020] which results in non-convex surrogates for linear FA, SPMA requires solving convex softmax classification problems in each iteration.

Contribution 3: In Section 2.4.2, we state the conditions under which Algorithm 1 converges to the neighbourhood of the optimal value function, and characterize the resulting linear convergence rate. Hence, for log-linear policies, Algorithm 1 matches the theoretical convergence of NPG with compatible function approximation [Agarwal et al., 2021a, Alfano and Rebeschini, 2022a, Yuan et al., 2023]. Our theoretical results are better than those in Vaswani et al. [2021] and Schulman et al. [2015a] which prove sublinear convergence to a stationary point for idealized variants of SPMA and TRPO respectively. In contrast to Kuba et al. [2022] which prove that the idealized variants of PPO and TRPO converge to the optimal policy asymptotically, we characterize the non-asymptotic convergence rate for Algorithm 1.

Contribution 4: We empirically evaluate SPMA across simple MDPs with tabular and linear parameterization, Atari games with a discrete action space and a neural policy parameterization with CNNs, and continuous control MuJoCo tasks with a continuous action space and a neural policy parameterization with MLPs. We demonstrate that SPMA has

consistently good performance – on Atari games SPMA achieves better results than both TRPO and PPO while matching or outperforming MDPO, whereas on MuJoCo tasks, SPMA outperforms PPO and achieves similar or better results than MDPO.

2.2 Problem Formulation

We consider an infinite-horizon discounted Markov decision process (MDP) [Puterman, 2014] defined by $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathcal{P}, r, \rho, \gamma \rangle$, where $\mathcal{S}$ and $\mathcal{A}$ represent the states and actions, $\mathcal{P} : \mathcal{S} \times \mathcal{A} \to \Delta_{\mathcal{S}}$ is the transition probability function, $r : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is the reward function, $\rho \in \Delta_{\mathcal{S}}$ is the initial state distribution, and $\gamma \in [0, 1)$ represents the discount factor. In this paper, we exclusively consider the setting where the number of states and actions is finite, but potentially large.

Given $s \in \mathcal{S}$, the policy π induces a probability distribution $\pi(.|s)$ over the actions. The action-value function $Q^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ induced by π is defined as $Q^\pi(s, a) := \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) | s_0 = s, a_0 = a]$ where $s_t \sim p(.|s_{t-1}, a_{t-1})$, and $a_t \sim \pi(.|s_t)$ for $t \geq 1$. The value function corresponding to Q^π starting from state s is defined as $V^\pi(s) := \mathbb{E}_{a \sim \pi(.|s)}[Q^\pi(s, a)]$ with $J(\pi) := V^\pi(\rho) := \mathbb{E}_{s \sim \rho}[V^\pi(s)]$ representing the expected discounted cumulative reward. Furthermore, the advantage function $A^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is defined as $A^\pi(s, a) := Q^\pi(s, a) - V^\pi(s)$. The policy also induces a discounted state-occupancy measure $d^\pi(s) := (1-\gamma) \sum_{t=0}^{\infty} \gamma^t \mathrm{Pr}^\pi[s_t = s | s_0 \sim \rho]$ over the states. The objective is to find an optimal policy π^* that maximizes the expected reward $J(\pi)$, i.e. $\pi^* = \arg\max_\pi J(\pi)$. As a special case, in the bandit setting, $|\mathcal{S}| = 1$, $|\mathcal{A}| = K$, $\gamma = 0$, and $J(\pi) = \langle \pi, r \rangle$, with K representing the number of arms.

2.3 Softmax Policy Mirror Ascent: Tabular Parametrization

Softmax policy mirror ascent (referred to as SPMA) represents the policy using the softmax function $h : \mathbb{R}^A \to \Delta_A$ i.e. $\pi(\cdot|s) = h(z(s, \cdot))$ s.t. for all $(s, a) \in \mathcal{S} \times \mathcal{A}$, $\pi(a|s) = \frac{\exp(z(s,a))}{\sum_{a'} \exp(z(s,a'))}$, where the *logits* z are SA-dimensional vectors and Δ_A is the A-dimensional simplex. We first focus on the *tabular parameterization* where the number of parameters scales with the number of states and actions, and aim to learn the logits corresponding to an optimal policy. With some abuse of notation, we use $J(z)$ to refer to $J(\pi)$ where $\pi(\cdot|s) = h(z(s, \cdot))$ and state the objective as: $\max_{z \in \mathbb{R}^{SA}} J(z)$.

As the name suggests, SPMA uses mirror ascent (MA) to maximize $J(z)$. For a differentiable, strictly convex mirror map Φ, MA [Beck and Teboulle, 2003, Bubeck et al., 2015] is an iterative algorithm whose update at iteration $t \in [T]$ can be stated in two equivalent ways:

$$\begin{aligned} \nabla \Phi(z_{t+1}) &= \nabla \Phi(z_t) + \eta \, \nabla_z J(z_t) \\ z_{t+1} &= \arg\max_{z \in \mathbb{R}^{SA}} \left[\langle z - z_t, \nabla_z J(z_t) \rangle - \frac{1}{\eta} D_\Phi(z, z_t) \right] \end{aligned} \tag{2.1}$$

where z_t is the logit at iteration t, η is the step-size and $D_\Phi(z, z') := \Phi(z) - \Phi(z') - \langle \nabla\Phi(z'), z - z' \rangle$ is the Bregman divergence between logits z and z' induced by the mirror map Φ. Hence, the MA update at iteration t can be interpreted as moving in the gradient direction $\nabla_z J(z_t)$ while staying "close" to the logit z_t, where the proximity between logits is measured according to the Bregman divergence and weighted by $\frac{1}{\eta}$.

2.3.1 Bandit Setting

It is instructive to first instantiate the SPMA update for the bandit setting where $J(\pi) = \langle \pi, r \rangle$. In this setting, $\nabla_z J(z) \in \mathbb{R}^A$ s.t. $[\nabla_z J(z)](a) = \pi(a)\,[r(a) - \langle \pi, r \rangle]$. Following Vaswani et al. [2021], we use the log-sum-exp mirror map i.e. $\phi(z) = \ln(\sum_a \exp(z(a))$. Since $[\nabla\phi(z)](a) = \frac{\exp(z(a))}{\sum_{a'} \exp(z(a'))} = [h(z)](a) = \pi(a)$, the logit and probability spaces are dual to each other, and the $\nabla\phi$ map can be used to move between these spaces. Given this, the SPMA update can be written as:

$$\begin{aligned} \pi_{t+1}(a) &= \pi_t(a)\,(1 + \eta\,[r(a) - \langle \pi, r \rangle]) \\ &= \pi_t(a)\,[1 + \eta \sum_{a' \neq a} \pi_t(a')\,\Delta(a, a')]\,, \end{aligned} \tag{2.2}$$

where $\Delta(a, a') := r(a) - r(a')$ represents the reward gap between arms a and a'. We first ensure that π_{t+1} is a valid probability distribution. Since $r(a) \in [0, 1]$ for all a, $\eta \leq 1$ is sufficient to guarantee that $\pi_{t+1}(a)$ is non-negative for every a. Moreover, $\sum_a \pi_{t+1}(a) = \sum_a \pi_t(a) + \eta \sum_a \pi_t(a)[r(a) - \langle \pi, r \rangle] = \sum_a \pi_t(a) = 1$. Hence, for $\eta \leq 1$, Equation (2.2) results in a valid update to the policy. The above update is related to the PROD algorithm [Cesa-Bianchi et al., 2007] used for the experts problem in the online learning literature. In contrast to SPMA which is derived from mirror ascent, PROD is derived using a linearization of the Hedge [Freund and Schapire, 1997] algorithm and requires explicit normalization to obtain probabilities.

2.3.2 MDP Setting

In order to extend SPMA to the MDP setting, we use a (state-wise) weighted log-sum-exp mirror map, i.e. for a logit $z \in \mathbb{R}^{SA}$, we define $\Phi(z) := \sum_s w(s)\,\phi(z(s, \cdot)) = \sum_s w(s) \ln(\sum_a \exp(z(s, a))$ where $w(s)$ are the per-state weights. Following the proof of Vaswani et al. [2024, Lemma 11], the resulting Bregman divergence is given as: $D_\Phi(z, z') = \sum_s w(s)\,\mathrm{KL}(\pi'(\cdot|s)||\pi(\cdot|s))$ where π and π' are the policies corresponding to logits z and z'. At iteration t of SPMA, we choose $w(s) = d^{\pi_t}(s)$ and use the policy gradient theorem [Sutton et al., 1999] to calculate

$[\nabla J(z_t)](s,a) = d^{\pi_t}(s)\,\pi_t(a|s)\,A^{\pi_t}(s,a)$. The resulting SPMA update is given as:

$$z_{t+1} = \arg\max_{z\in\mathbb{R}^{SA}} \sum_s d^{\pi_t}(s)\Big[\langle \pi_t(\cdot|s)\,A^{\pi_t}(s,\cdot),\, z(s,\cdot)\rangle - \frac{1}{\eta}\,\mathrm{KL}(\pi_t(\cdot|s)\,||\,h(z(s,\cdot))\Big].$$

Since the above maximization decomposes over the states, we can write the per-state update for each $s \in \mathcal{S}$ in terms of $\pi_{t+1}(\cdot|s) = h(z_{t+1}(s,\cdot))$ as follows:

$$\pi_{t+1}(a|s) = \pi_t(a|s)\,(1 + \eta A^{\pi_t}(s,a))\,. \tag{2.3}$$

Similar to the bandit case, since $r(s,a) \in [0,1]$, $\pi_{t+1}(a|s)$ is non-negative for $\eta \le 1-\gamma$. Since $\sum_a \pi_t(a|s)A^{\pi_t}(s,a) = 0$, $\sum \pi_{t+1}(a|s) = 1$, and hence Equation (2.3) results in a valid policy update.

In order to compare the SPMA update to existing methods, note that for the tabular parameterization, natural policy gradient (NPG) update [Kakade, 2001] is the same as policy mirror ascent [Lan, 2023, Johnson et al., 2023, Xiao, 2022] and is given by: $\pi_{t+1}(a|s) \propto \pi_t(a|s)\,\exp(\eta\,A^{\pi_t}(s,a))$. In contrast to NPG, the SPMA update in Equation (2.3) is linear in both η and $A^{\pi_t}(s,a)$ and does not require an explicit normalization across actions to ensure valid probability distributions. On the other hand, softmax policy gradient (SPG) [Agarwal et al., 2021a, Mei et al., 2020] corresponds to choosing the mirror map ϕ in Equation (2.1) to be the Euclidean norm and has the following update: $z_{t+1}(s,a) = z_t(s,a) + \eta\,\pi_t(a|s)A^{\pi_t}(s,a)$. Compared to SPG that uses the softmax policy gradient to update the logits, SPMA uses the softmax policy gradient to directly update the probabilities. As we demonstrate in the next section, this desirable property enables SPMA to achieve faster rates than SPG.

2.3.3 Theoretical Results

In this section, we prove convergence guarantees for SPMA in the multi-armed bandit and tabular MDP settings. We first establish linear convergence for SPMA for multi-armed bandits for any constant $\eta \le 1$.

Theorem 1. *The* SPMA *update in Equation* (2.2) *with (i) a constant step-size* $\eta \le 1$*, and (ii) uniform initialization i.e.* $\pi_0(a) = \frac{1}{K}$ *for all* a *converges as:*

$$r(a^*) - \langle \pi_T, r\rangle \le \left(1 - \frac{1}{K}\right)\exp\left(\frac{-\eta\,\Delta_{\min}\,T}{K}\right),$$

where T *is the number of iterations,* a^* *is the optimal arm i.e.* $a^* = \arg\max_a r(a)$ *and* $\Delta_{\min} := \min_{a\neq a^*} \Delta(a^*,a) = r(a^*) - r(a)$ *is the gap.*

The above theorem (proved in Section A.1) shows that for multi-armed bandit problems, SPMA can achieve linear convergence to the optimal arm, and the resulting rate depends on both the gap and the number of arms. In Section A.1.1, we prove that SPMA with specific

gap-dependent step-sizes can achieve a global super-linear convergence rate for multi-armed bandits. To the best of our knowledge, these are the first global super-linear rates for PG methods on multi-armed bandit problems.

In the next theorem, we extend the linear convergence result to tabular MDPs and prove that when given access to the exact policy gradients, SPMA results in linear convergence to the optimal value function for any sufficiently small constant step-size.

Theorem 2. *Using the* SPMA *update in Equation* (2.3) *with a step-size* $\eta < \min\left\{1-\gamma, \frac{1}{C_t(1-\gamma)}\right\}$ *converges as:*

$$\left\|V^{\pi^*} - V^{\pi_T}\right\|_\infty \leq \left(\prod_{t=0}^{T-1} \alpha_t\right) \left\|V^{\pi^*} - V^{\pi_0}\right\|_\infty ,$$

where $\alpha_t := (1 - \eta\, C_t\,(1-\gamma))$, $C_t := \min_s\{\pi_t(\tilde{a}_t(s)|s)\,\Delta^t(s)\}$, $\tilde{a}_t(s) := \arg\max_a Q^{\pi_t}(s,a)$ *and* $\Delta^t(s) := \max_a Q^{\pi_t}(s,a) - \max_{a\neq\tilde{a}} Q^{\pi_t}(s,a)$.

For ease of exposition, the above theorem considers $\tilde{a}_t(s)$ to be the unique action maximizing $Q^{\pi_t}(s,\cdot)$ for every state s and policy π_t. In Section A.2, we extend this to include multiple optimal actions with a minor change in the definition of the gap. For rewards in $(0,1)$, $C_t\,(1-\gamma)$ is in $(0,1)$ and depends on the initialization. If $C := \min_{t\in[T]} C_t$, then the above implies that when $T \in O\left(\frac{1}{\eta\, C(1-\gamma)}\, \ln(1/\epsilon)\right)$, SPMA guarantees that $V^{\pi_T}(s) \geq V^*(s) - \epsilon$ for all $s \in \mathcal{S}$. Note that in order to establish linear convergence, it is crucial to ensure that C is nonzero. Proving this property theoretically is challenging, as reflected in related work. In particular, the best-known linear convergence rates for NPG with a constant step size also depend on this same constant (see, for example, Theorem 5.4 in [Liu et al., 2024] and Lemma 10 in [Mei et al., 2021a]). Consequently, in Section A.3.3, we empirically verify that C is lower-bounded by a positive constant.

In order to put the above convergence result in context, note that SPG with a constant step-size results in a $\Theta\,(1/\epsilon)$ convergence [Mei et al., 2020]. Recently, Chen et al. [2023] proved that constant step-size SPG with Nesterov acceleration can obtain an $O(1/\sqrt{\epsilon})$ convergence. In contrast, the above theorem demonstrates that by choosing the appropriate mirror map, constant step-size SPMA can achieve a faster $O(\log(1/\epsilon))$ rate of convergence. On the other hand, Liu et al. [2024], Lu et al. [2024] prove that SPG with adaptive step-sizes can also result in linear convergence. However, the resulting rate depends on the distribution mismatch ratio $\left\|\frac{d^{\pi^*}}{\rho}\right\|_\infty$ that can be exponentially large in the size of the state space [Li et al., 2021]. In contrast, the convergence result in Theorem 2 has no such dependence. The linear convergence rate in Theorem 2 matches that of NPG with a constant step-size [Liu et al., 2024] and compared to Liu et al. [2024, Theorem 5.4], it results in a better dependence (exponential vs polynomial) on the gap $\Delta^t(s)$. Finally, we note that for the tabular parameterization, a variant of TRPO has been shown to achieve $O(1/\epsilon^2)$ convergence to the optimal policy [Shani et al., 2020].

In the next section, we extend SPMA to exploit function approximation to handle large state-action spaces.

2.4 Handling Function Approximation

Handling large MDPs requires function approximation (FA) techniques to share information between states and actions. For example, given a set of state-action features $\mathbf{X} \in \mathbb{R}^{SA \times d}$ where $d << SA$, the *log-linear policy parameterization* [Agarwal et al., 2021a, Alfano and Rebeschini, 2022a, Yuan et al., 2023] considers policies of the form: $\pi(a|s) = \frac{\exp(\langle \mathbf{X}(s,a), \theta \rangle)}{\sum_{a'} \exp(\langle \mathbf{X}(s,a'), \theta \rangle)}$ where $\theta \in \mathbb{R}^d$ is the parameter to be learned. Hence, the log-linear policy parameterization can handle large state-action spaces while learning a compressed d-dimensional representation. We interpret the log-linear policy parameterization as a constraint in the space of logits.

Algorithm 1: SPMA with function approximation

Input: θ_0 (parameters for the initial policy π_0), f_θ (function approximation), T (number of outer-loop), m (number of inner-loops), η (outer-loop step-size), ζ (inner-loop step-size)

for $t \leftarrow 0$ **to** $T-1$ **do**

1. Interact with the environment using π_t and form the surrogate function $\ell_t(\theta)$ in Equation (2.5)
2. Initialize inner-loop: $\omega_0 = \theta_t$
   for $k \leftarrow 0$ **to** $m-1$ **do**
   $\omega_{k+1} = \omega_k - \zeta \nabla_\omega \ell_t(\omega_k)$
   end
3. $\theta_{t+1} = \omega_m$
4. Update $\pi_{t+1}(\cdot|s) = h(f_{\theta_{t+1}}(s, .))$

end

Return θ_T

Specifically, the logits z are constrained to lie in the set $\mathcal{Z} = \{z \in \mathbb{R}^{SA} | \exists \theta \text{ s.t. } z = \mathbf{X}\theta\}$, meaning that the logits are required to be realizable by the linear model with features $\mathbf{X}$. We define Π as the corresponding set of feasible policies, i.e. $\Pi = \{\pi | \forall s \in \mathcal{S}, \pi(\cdot|s) = h(z(s,\cdot)) \text{ s.t. } z \in \mathcal{Z}\}$. Hence, the policies in Π are constrained to be log-linear. Note that, as in the case of log-linear policies, Π can be a non-convex set, even when $\mathcal{Z}$ is convex. For general *energy-based or neural policies* [Haarnoja et al., 2017, Agarwal et al., 2021a], $\pi(a|s) \propto \exp(f_\theta(s,a))$ where $f_\theta : \mathbb{R}^{SA} \to \mathbb{R}$ is a complex, non-linear model. In this case, the logits are constrained to lie in the set: $\mathcal{Z} = \{z \in \mathbb{R}^{SA} | \exists \theta \text{ s.t. } z(s,a) = f_\theta(s,a)\}$.

The above interpretation allows us to extend SPMA to the FA setting. Specifically, we use the same mirror ascent update as in Equation (2.1) with an additional projection step onto the feasible set $\mathcal{Z}$. Specifically, we define $z_{t+1/2}$ s.t. $\nabla\Phi(z_{t+1/2}) = \nabla\Phi(z_t) + \eta \nabla_z J(z_t)$ and compute $z_{t+1} = \arg\min_{z \in \mathcal{Z}} D_\Phi(z, z_{t+1/2})$. This step denotes the Bregman projection of $z_{t+1/2}$ onto $\mathcal{Z}$, i.e. we seek to find the closest (according to the Bregman divergence) realizable point (in $\mathcal{Z}$) to the "ideal" point $z_{t+1/2}$ which corresponds to using the tabular parameterization.

Following Vaswani et al. [2021], Lavington et al. [2023], we convert the above projection problem into an unconstrained minimization problem where $\forall(s,a)$, $z_{t+1}(s,a) = f_{\theta_{t+1}}(s,a)$, $z_\theta(s,a) := f_\theta(s,a) \in \mathcal{Z}$, $\theta_{t+1} = \arg\min_\theta D_\Phi(z_\theta, z_{t+1/2})$ i.e. we aim to find the parameter θ that realizes the point $z_\theta \in \mathcal{Z}$ which is closest to $z_{t+1/2}$. Following Section 2.3.2, using the log-sum-exp mirror map weighted by $d^{\pi_t}(s)$ at iteration t results in the following optimization problem $\theta_{t+1} = \arg\min_\theta \tilde{\ell}_t(\theta)$ where,

$$\begin{aligned}\tilde{\ell}_t(\theta) &:= \sum_s d^{\pi_t}(s)\,\mathrm{KL}(\pi_{t+1/2}(\cdot|s)\,||\,\pi_\theta(\cdot|s)) \\ &= \mathop{\mathbb{E}}_{s\sim d^{\pi_t}} \mathcal{H}\left(h(f_{\theta_t}(s,\cdot))(1+\eta A^{\pi_t}(s,\cdot)), h(f_\theta(s,\cdot))\right) + C_t\,.\end{aligned} \tag{2.4}$$

Here, $\mathcal{H}(p,q) := -\mathbb{E}_p[\ln(q)] = -\sum_a p(a)\ln(q(a))$ is the cross-entropy between distributions p and q and C_t is a constant with respect to θ. We refer to $\tilde{\ell}_t(\theta)$ as the *ideal surrogate*. Minimizing this surrogate requires calculating the expectation over the states sampled according to π_t. In order to have a practical algorithm, we can run trajectories τ starting from the initial state distribution ρ, following the policy π_t and thus sampling from the d^{π_t} distribution (see Agarwal et al. [2021a, Algorithm 3] for the detailed procedure). Given these samples, we form the surrogate $\ell_t(\theta)$ defined as:

$$\sum_{s\sim\tau} \mathrm{KL}\left(h(f_{\theta_t}(s,\cdot))(1+\eta A^{\pi_t}(s,\cdot))\,||\,h(f_\theta(s,\cdot))\right)\,. \tag{2.5}$$

Note that $\mathbb{E}[\ell_t(\theta)] = \tilde{\ell}_t(\theta)$ where the expectation is w.r.t. to d^{π_t}. We use m steps of (stochastic) gradient descent to approximately minimize $\ell_t(\theta)$. Putting everything together, the algorithm incorporating general FA is presented in Algorithm 1.

Log-linear Policy Parameterization: For this special case, the problem in Equation (2.4) is equivalent to a weighted (according to $d^{\pi_t}(s)$) multi-class classification for each state. The per-state problem corresponds to a softmax classification into A classes using a linear model with features $\mathbf{X}$ and soft labels equal to $\pi_{t+1/2}(\cdot|s)$. Computing θ_{t+1} thus involves minimizing a smooth, convex function.

In the next section, we compare Algorithm 1 to existing approaches that incorporate FA.

2.4.1 Comparison to Existing Approaches

Comparison to `NPG`: A principled extension of `NPG` to handle FA is via the compatible function approximation [Kakade, 2001, Agarwal et al., 2021a]. An example of such an algorithm, `Q-NPG` involves solving a quadratic surrogate at each iteration t: $\hat{\omega}_t = \min_\omega \sum_s d^{\pi_t}(s)\sum_a \pi_t(a|s)\left(f_\omega(s,a) - Q^{\pi_t}(s,a)\right)^2$. The policy parameters are updated using $\hat{\omega}_t$ which corresponds to the natural gradient direction. While this approach results in theoretical guarantees (see Section 2.4.2 for details); for a general parameterization, the

resulting algorithm involves changing the representation of the critic at every iteration. Consequently, solving the surrogate is expensive, limiting the practicality of the method.

Comparison to MDPO: A more practical extension of NPG is mirror descent policy optimization [Tomar et al., 2020] (MDPO). Similar to SPMA, MDPO can be interpreted as projected (onto the feasible set of policies) mirror ascent in the space of probabilities [Vaswani et al., 2021]. The resulting surrogate (as a function of the policy parameters) is given by: $\sum_s d^{\pi_t}(s)\,\mathrm{KL}\left(\pi_\theta(\cdot|s)\,||\,h\left(f_{\theta_t}(s,\cdot)\,\exp(\eta\,Q^{\pi_t}(s,\cdot))\right)\right)$. Unlike the surrogate in Equation (2.4), the MDPO surrogate is non-convex even when using a tabular softmax parameterization for the policy, and consequently does not have any theoretical guarantees. However, MDPO results in good empirical performance, and we compare to it in Section 2.5.

Comparison to TRPO: As explained in Vaswani et al. [2021], the surrogate in Equation (2.4) is closely related to TRPO. In particular, the TRPO update consists of solving the following optimization problem: $\sum_s d^{\pi_t}(s)\sum_a \pi_t(a|s)A^{\pi_t}(s,a)\,\frac{\pi_\theta(a|s)}{\pi_{\theta_t}(a|s)}$, such that $\mathbb{E}_{s\sim d^{\pi_t}}\left[\mathrm{KL}(\pi_t(\cdot|s)\,||\,\pi_\theta(\cdot|s))\right] \leq \delta$. SPMA (i) uses instead $\sum_s d^{\pi_t}(s)\sum_a \pi_t(a|s)A^{\pi_t}(s,a)\,\log\frac{\pi_\theta(a|s)}{\pi_{\theta_t}(a|s)}$, i.e. the logarithm of the importance sampling ratio, making the resulting update more stable [Vaswani et al., 2021] and (ii) enforces the proximity between policies via a regularization rather than a constraint. Enforcing the trust-region constraint in TRPO requires additional hyper-parameters, code-level optimizations and computation [Engstrom et al., 2019]. In contrast, SPMA is more computationally efficient and simpler to implement in practice. In the next section, we study the theoretical properties of Algorithm 1.

2.4.2 Theoretical Guarantee

For rewards in $[0,1]$ and for a general policy parameterization, Vaswani et al. [2021] prove that, for $\eta \leq 1-\gamma$, Algorithm 1 results in monotonic improvement, i.e. $J(\pi_{t+1}) \geq J(\pi_t)$ and hence converges to a stationary point at an $O(1/\epsilon)$ rate. Since J is non-convex and can have multiple stationary points, the result in Vaswani et al. [2021] does not provide sufficient evidence for the good empirical performance of Algorithm 1. In this section, we prove that, under reasonable assumptions similar to existing works, Algorithm 1 can converge to the neighbourhood of the optimal value function at a linear rate. The size of the neighbourhood is determined by various practical factors such as sampling, inexact optimization, and bias due to the FA. In order to state our result, we first state and justify our assumptions.

Recall that in order to have a practical algorithm, we minimize $\ell_t(\theta)$ obtained by sampling from d^{π_t}.

Assumption 1. *Excess Risk: For all iterations t of Algorithm 1,* $|\tilde{\ell}_t(\theta_{t+1}) - \min \tilde{\ell}_t(\theta)| \leq \epsilon_{stat}$.

The above assumption quantifies the excess risk incurred by minimizing a finite sampled "dataset" of states as compared to minimizing over the population loss $\tilde{\ell}_t(\theta)$. This is a standard assumption in the literature analyzing the convergence of policy gradient methods

with FA [Agarwal et al., 2021a, Alfano and Rebeschini, 2022a, Yuan et al., 2023]. If n is the number of samples and the surrogate is minimized using (stochastic) gradient descent, using the standard generalization results [Lei and Ying, 2021, Nikolakakis et al., 2022], we expect $\epsilon_{\text{stat}} = O(1/n)$ for the log-linear parameterization. For example, using m iterations of SGD would result in $\epsilon_{\text{stat}} = O(1/n + 1/m)$ [Lei and Ying, 2021, Theorem 6]. For a general parameterization, where the surrogate might be non-convex, the excess risk can be bounded up to the optimization error [Nikolakakis et al., 2022]. Under the appropriate technical assumptions, $\ell_t(\theta)$ can been shown to satisfy the Polyak-Lojasiewicz condition [Liu et al., 2022] implying that the optimization error for (stochastic) gradient descent can be made arbitrarily small. The next assumption quantifies the bias incurred because of a policy parameterization with limited expressive power compared to using the tabular parameterization.

Assumption 2. *Bias: For all iterations t of Algorithm 1,* $\min_\theta \tilde{\ell}_t(\theta) \leq \epsilon_{bias}$.

The above assumption captures the flexibility of the model class being used in the policy parameterization. For a tabular parameterization where the number of parameters scales as SA, $\epsilon_{\text{bias}} = 0$, whereas for the log-linear parameterization, ϵ_{bias} depends on the expressivity of the features. The final assumption is concerned with exploration and indicates that the initial state distribution has full support implying that the method does not require explicit exploration.

Assumption 3. *Exploration:* $\forall s \in \mathcal{S}$, $\rho(s) \geq \rho_{\min} > 0$.

The above assumption is standard in the literature [Agarwal et al., 2021a, Xiao, 2022] and helps isolate and study the optimization properties of PG methods. We prove the following theorem in Section A.2.

Theorem 3. *Under Assumption 1-3,* SPMA *with* $\eta < \min\left\{1-\gamma, \frac{1}{C_t(1-\gamma)}\right\}$ *converges as,*

$$
\begin{aligned}
&J(\pi^*) - J(\pi_T) \\
&\leq \left(\prod_{t=0}^{T-1} \alpha_t\right)(J(\pi^*) - J(\pi_0)) + \beta \sum_{t=0}^{T-1} \prod_{i=t+1}^{T-1} \alpha_i \,,
\end{aligned}
$$

where $\beta = \frac{\sqrt{2}}{(1-\gamma)^2 \rho_{\min}} \sqrt{\epsilon_{stat} + \epsilon_{bias}}$ *and* α_t *has the same definition as in Theorem 2.*

The above theorem shows that Algorithm 1 converges linearly to the neighbourhood of the optimal value function. Furthermore, for the log-linear parameterization, the size of the neighbourhood can be bounded explicitly. For example, if the logits $z_{t+1/2}$ for every t lie in the span of the features, $\epsilon_{\text{bias}} = 0$ (this is similar to the linear Bellman completeness condition used in the analysis of value-based methods [Munos, 2005]) and $\epsilon_{\text{stat}} = O(1/n + 1/m)$. By

using more samples and with more (S)GD iterations, the size of the neighbourhood can be made arbitrarily small. Except for the neighbourhood term, the above convergence result is similar to that for the tabular setting in Theorem 2. The other difference is that the result in the tabular setting holds in the ℓ_∞ norm and thus holds for all states, whereas the result in Theorem 3 only holds for a fixed starting state distribution ρ. In practice, A^{π_t} is typically estimated via a critic. To account for this, we generalize the proof of Theorem 3 in Section A.2, and prove that Algorithm 1 converges linearly to a neighbourhood that depends on an additional term proportional to the critic error.

We now compare to the existing theoretical results for PG methods with FA. For the log-linear policy parameterization, `Q-NPG` and its variants have been shown to achieve linear convergence to the neighbourhood of the optimal value function [Agarwal et al., 2021a, Alfano and Rebeschini, 2022a, Yuan et al., 2023]. The size of the neighbourhood depends on similar quantities as Theorem 3. Finally, we note that while an ideal, impractical variant of `TRPO` has a monotonic improvement guarantee similar to Algorithm 1 [Schulman et al., 2015a], it does not have convergence guarantees comparable to Theorem 3.

2.5 Empirical Evaluation

We evaluate `SPMA`[1] on three types of problems: (i) tabular MDPs with access to exact policy gradients, (ii) MDPs with continuous states but discrete actions, using inexact policy gradients, and (iii) MDPs with continuous state-actions spaces and inexact gradients. For tabular MDPs, we use the tabular parameterization and compare `SPMA` against `NPG` and constant step-size `SPG` [Mei et al., 2020]. For these environments, we also consider log-linear policies and compare `SPMA` to `MDPO` and `SPG`. For non-tabular environments, we consider `PPO`, `TRPO` and `MDPO` as baselines. We consider two variants of `TRPO` – `TRPO-constrained`, the standard optimized variant in Raffin et al. [2021] and `TRPO-regularized`, the original regularized variant. `TRPO-constrained` is able to effectively enforce the trust region constraint using conjugate gradient, but introduces additional hyper-parameters, requires code-level optimization techniques and is computationally expensive. On the other hand, `TRPO-regularized` is significantly more efficient and theoretically principled [Lazić et al., 2021], and is similar to `SPMA`'s objective (see Section 2.4.1). For details regarding the hyper-parameters of all methods for each environment, refer to Sections A.3 and A.4.

Tabular MDP Results: We present the results in Section A.3, and summarize the key findings here. We observe that `SPMA` and `NPG` achieve comparable performance, both consistently outperforming `SPG` (Figure A.1). However, in the linear FA setting, `SPMA` demonstrates superior performance compared to `MDPO` in the CliffWorld environment [Sutton and Barto,

[1]The code is available at `http://github.com/reza-asad/SPMA`.

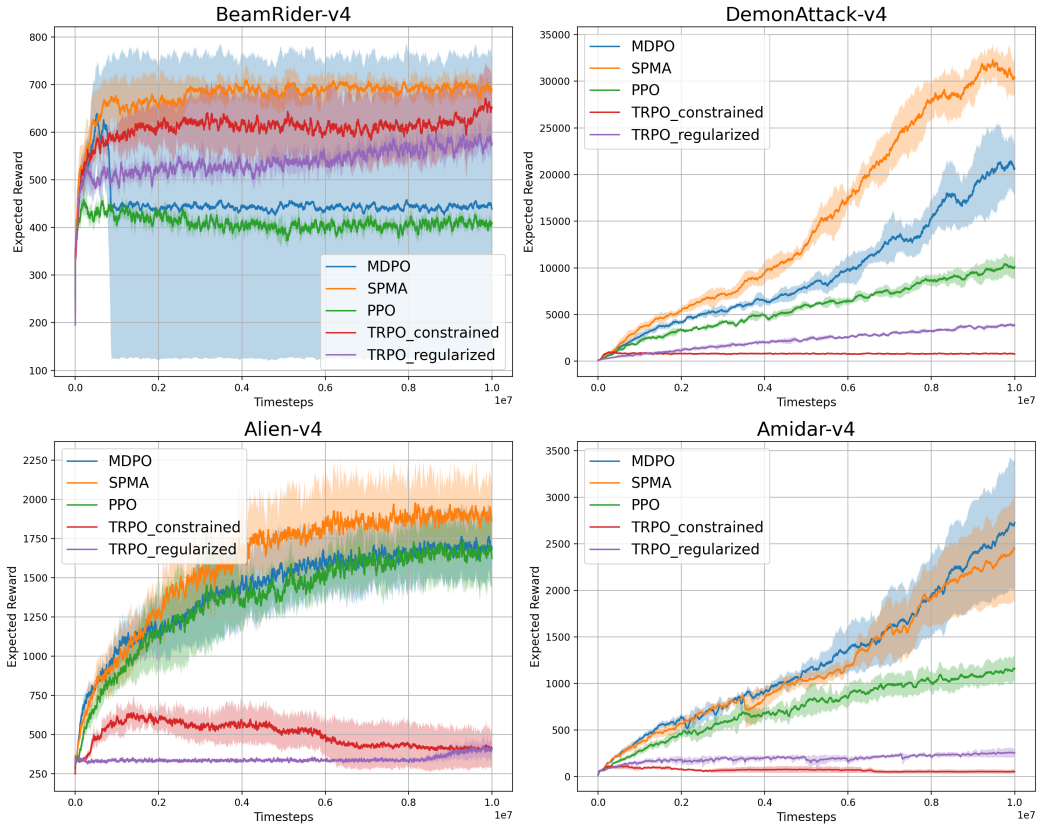


Figure 2.1: On Atari games, where a CNN-based actor network is employed, SPMA matches or surpasses MDPO and outperforms PPO, as well as both the constrained and regularized versions of TRPO.

2018] (Figure A.3), while performing similarly in the Frozen Lake environment [Brockman, 2016] (Figure A.4). In both environments, SPMA and MDPO consistently outperform SPG.

In the remainder of this section, we focus on the non-tabular settings with inexact policy gradients. For these experiments, we follow the protocol of Tomar et al. [2020], using 5 seeds and reporting the average results along with their 95% confidence intervals. Additionally, we employ the actor-critic architecture, policy parameterization, and GAE [Schulman et al., 2015b] (to estimate the advantage function) from stable baselines [Raffin et al., 2021]. We emphasize that, in contrast to prior work, we do not make ad-hoc adjustments to SPMA (i.e., the actor). To set the step-size η in Algorithm 1, we perform a grid search over five values (fixed across all experiments) and set the inner loop step-size ζ using Armijo line search [Armijo, 1966].

Atari and Mujoco Results: We evaluate the performance of SPMA compared to the baselines across various Atari 2600 games [Bellemare et al., 2013] and MuJoCo [Todorov et al., 2012] control tasks from OpenAI Gym [Brockman, 2016]. The observation space for Atari games consists of a 210×160×3 image, representing the current state of the game. The

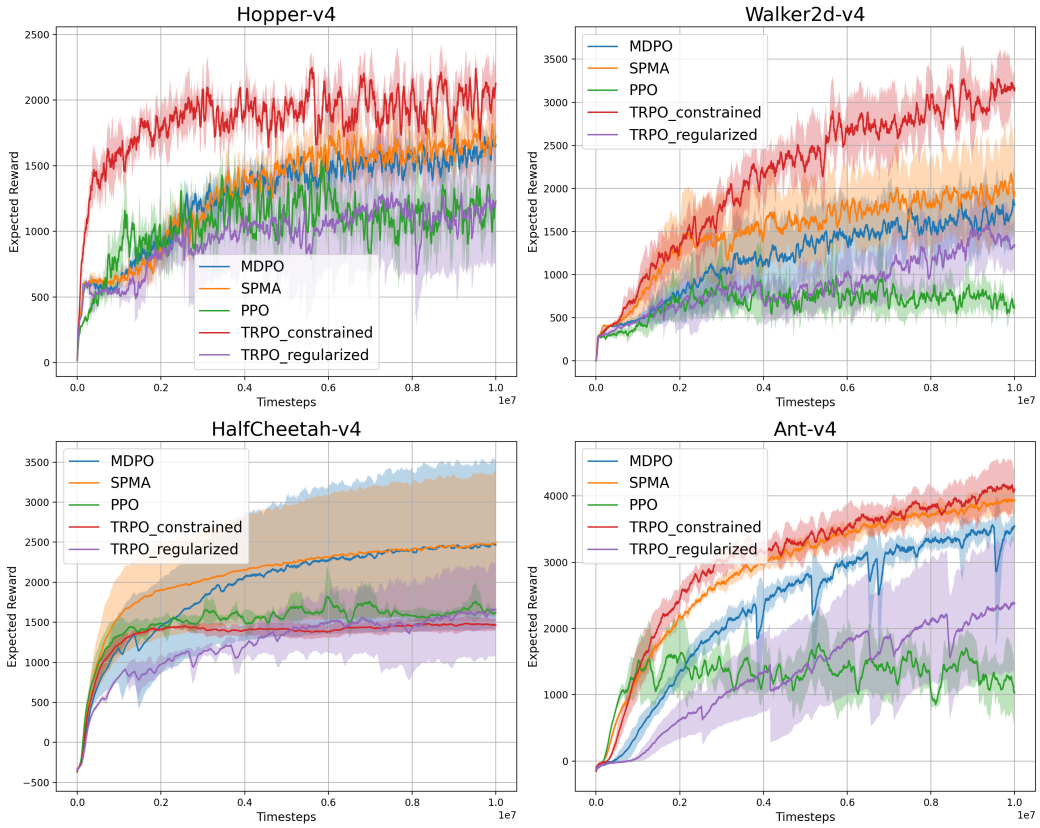


Figure 2.2: On MuJoCo control tasks, where a two-layer MLP actor network is used, `SPMA` matches or outperforms `MDPO` while consistently outperforming `PPO` and regularized `TRPO`. In contrast to the results on Atari games, with a shallow MLP, `TRPO-constrained` outperforms all methods.

action space in these environments is discrete, whereas in MuJoCo, it is continuous and by default represented by a diagonal Gaussian distribution in Raffin et al. [2021]. Additionally, the actor-critic network for Atari uses a CNN as a feature extractor, while MuJoCo employs an MLP.

Comparing the results in Fig. 2.1 and 2.2, our key observations are as follows: i) `SPMA` consistently outperforms or matches `MDPO` and `PPO` across all environments; ii) although `TRPO-constrained` achieves superior performance on MuJoCo, its performance degrades considerably on Atari games. We conjecture that the conjugate gradient algorithm in `TRPO-constrained` performs poorly when the actor network is a CNN rather than a two-layer MLP; iii) `TRPO-regularized`, which has a similar objective as `SPMA` (see Section 2.4.1 for a comparison) does not perform as well on MuJoCo and has considerably worse performance on Atari. Hence, we observe that replacing the sampling ratio by its log can result in substantial empirical gains. This behaviour has also been observed for `PPO` Vaswani et al. [2021]. Overall, our experiments demonstrate that, despite being theoretically grounded, `SPMA` exhibits strong empirical performance across various environments without relying on ad-hoc adjustments.

2.6 Discussion

We developed SPMA, a PG method that corresponds to mirror ascent in the dual space of logits. We believe that our paper bridges the gap between theoretical PG methods and practical objectives by presenting a method that offers strong theoretical convergence guarantees while delivering competitive practical performance (compared to PPO, TRPO, MDPO), without relying on additional heuristics or algorithmic modifications. In the future, we aim to develop techniques for adaptively tuning the step-size and avoiding expensive grid-searches.

Towards off-policy actor-critic methods. The practical implementation of SPMA uses a learned critic to estimate advantages, while our analysis focuses on the policy update and accounts for critic error without explicitly analyzing the critic-learning procedure. This chapter also focuses on the on-policy setting, whereas reusing data collected by earlier policies can improve sample efficiency [Haarnoja et al., 2018b]. These considerations motivate extending the study to the joint design and analysis of actor and critic updates, together with off-policy data reuse. For example, an off-policy actor-critic algorithm can use a replay buffer of transitions collected by earlier policies to train its critic and update its current actor. However, the algorithm's behavior depends on how policy evaluation and policy improvement are combined. Soft Actor-Critic provides a successful example in continuous control [Haarnoja et al., 2018a], yet its discrete-action counterpart performs poorly on Atari [Christodoulou, 2019, Zhou et al., 2022]. The next chapter investigates this gap by identifying the bottleneck in discrete Soft Actor-Critic and developing a modular off-policy actor-critic framework for discrete actions that extends SPMA to the off-policy setting.

Chapter 3

Dissecting Discrete Soft Actor-Critic: Limitations and Principled Alternatives

3.1 Introduction

Soft Actor-Critic (SAC) [Haarnoja et al., 2018a] is a widely used off-policy reinforcement learning algorithm for continuous control, combining entropy-regularized critic learning with policy optimization. It has achieved strong empirical performance across simulated control benchmarks and real-world robotic systems [Haarnoja et al., 2018b], and has inspired numerous extensions that further improve its effectiveness [Zhu et al., 2024, Chen et al., 2021, Hiraoka et al., 2021, Bhatt et al., 2019, Kuznetsov et al., 2020, Ishfaq et al., 2025]. Motivated by this success, prior work has sought to adapt SAC to off-policy discrete-action settings [Christodoulou, 2019], which are central to domains such as Atari. However, despite being a natural extension, the resulting discrete variant of SAC (DSAC) exhibits weak empirical performance on standard benchmarks.

To address this limitation, follow-up work has proposed alternative discrete-action extensions of SAC [Xu et al., 2021, Zhou et al., 2022]. However, these methods typically introduce multiple interacting components that complicate implementation and can limit practical performance. For example, SD-SAC [Zhou et al., 2022] relies on ad-hoc actor regularization and double-averaged Q-clipping to achieve good performance on Atari. Other lines of work propose alternative off-policy actor-critic designs for the discrete-action setting, including off-policy variants of PPO [Queeney et al., 2021, Meng et al., 2023, Gan et al., 2024] and the GreedyAC algorithm [Neumann et al., 2018]. While these latter approaches offer complementary perspectives, they have not been evaluated on large-scale benchmarks such as Atari.

Motivated by the strong performance of continuous SAC and the limited success of its discrete counterpart, DSAC, we seek to address the following central question:

Can we isolate the bottleneck limiting DSAC*'s performance and identify design choices for building principled and efficient off-policy discrete-action actor–critic methods?*

To achieve our objective, we first investigate the poor Atari performance of DSAC [Christodoulou, 2019]. Prior work attributes this weakness to the use of a fixed target entropy [Xu et al., 2021], unstable coupling between policy and critic updates, and Q-value underestimation bias [Zhou et al., 2022]. Consequently, existing remedies adapt the entropy target [Xu et al., 2021] or add an entropy penalty along with clipped double-average Q-learning [Zhou et al., 2022]. While these modifications can stabilize training, they introduce extra hyper-parameters and algorithmic complexity. For example, the Atari experiments in Zhou et al. [2022] replace the automatic entropy-tuning mechanism of the original DSAC with a fixed entropy coefficient, introducing an additional tuning burden. Other related work includes Neumann et al. [2025] that modifies the DSAC objective and proposes surrogates with explicit KL regularization, but does not evaluate them on the standard Atari benchmark. Importantly, none of these methods admit theoretical guarantees, even in the tabular setting.

Contribution 1: Our extensive ablation study shows that critic entropy substantially impacts the empirical performance of DSAC. In particular, disabling the entropy regularization in the critic update (i.e., using the standard Bellman operator for policy evaluation) while keeping all other components (entropy-regularized soft actor update, automatic entropy tuning) fixed yields a stable variant of DSAC. This variant does not require double Q-learning or additional hyperparameters, and significantly outperforms the default DSAC across Atari games (see Figure 3.1). Furthermore, in Section 3.5 (Conclusion 1), we show that this method also outperforms prior DSAC variants [Xu et al., 2021, Zhou et al., 2022] on Atari, including hard-exploration games.

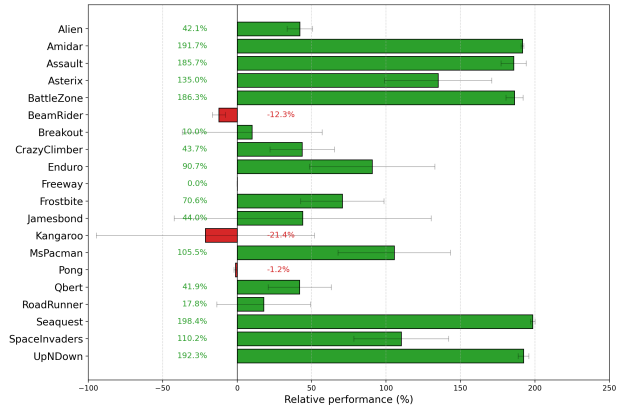


Figure 3.1: Relative performance of DSAC *without critic entropy* (our contribution) compared to the default DSAC across 20 Atari games, averaged over 5 seeds with 95% confidence intervals. Removing the critic entropy leads to substantial performance gains.

In order to systematically design and study algorithms related to DSAC, it is necessary to develop a more general approach. Prior works [Vieillard et al., 2020a, Xiao, 2022] have proposed such actor-critic frameworks in the discrete-action setting. While Vieillard et al. [2020a] show that the discrete extension of SAC falls within their framework, our proposed DSAC variant can not be captured by it. In particular, in Vieillard et al. [2020a], the actor

and critic entropy are closely coupled, and the framework cannot support using entropy regularization asymmetrically (e.g. entropy regularization for actor, but not for the critic). On the other hand, Xiao [2022] propose a policy gradient framework and analyze the actor objective in DSAC. However, they assume that Q is estimated via a black-box procedure and do not instantiate the critic update.

Contribution 2: In Section 3.3, we develop a general off-policy actor-critic framework for the discrete-action setting. Importantly, our framework decouples critic and actor entropy, enabling greater flexibility. The modular structure of our actor update gives rise to a family of objectives, including two novel ones, and subsumes DSAC as a special case. In particular, in the policy evaluation step for the critic, we use a look-ahead target formed by either the soft *critic entropy*-regularized or the standard hard Bellman operator. The policy optimization step for the actor consists of two stages: (i) computing an intermediate policy and (ii) projecting it onto the class of realizable policies while simultaneously maximizing a proximal entropy regularization term (referred to as the *actor entropy*). For (i), we consider two methods —NPG [Kakade, 2001] due to its widespread use, and SPMA [Asad et al., 2025b] because it can avoid normalization across actions and has demonstrated good empirical performance in the on-policy setting. For (ii), following [Chan et al., 2022], we study both forward and reverse KL divergences for the projection step to understand the effect of this choice within the proposed framework.

Contribution 3: Unlike prior work that extends SAC to discrete-action settings or studies its weak empirical performance [Christodoulou, 2019, Xu et al., 2021, Zhou et al., 2022], our framework has a strong theoretical foundation. In Section 3.4, we analyze the proposed methods in the tabular setting under asymmetric actor and critic entropy and establish the first convergence guarantees for such methods. Moreover, when the actor and critic entropy is coupled, we recover the convergence rates in Vieillard et al. [2020a]. More concretely, in Theorem 4, we reduce the problem of analyzing the sub-optimality in the entropy-regularized value function to (i) bounding the policy evaluation error for the critic and (ii) bounding the regret for an online convex optimization problem related to the policy update step for the actor. Our modular framework can be used to analyze different combinations of the actor and critic. For example, in Corollary 13, we leverage our framework to provide a theoretical guarantee for the NPG update with actor entropy and m steps of the hard Bellman operator (i.e., no critic entropy) for policy evaluation. This corresponds to one of the variants that we later evaluate empirically in the function approximation setting.

Contribution 4: In Section 3.5, we empirically evaluate the objectives instantiated using our actor-critic framework. Our results reveal three key findings. First, similar to DSAC, the objectives derived from our framework benefit from the hard Bellman operator for policy evaluation, resulting in improved performance. Second, unlike DSAC, the performance of these objectives, including two novel ones, remains robust even without entropy regularization or

explicit exploration. Third, using the forward KL divergence when projecting the intermediate policy onto the class of realizable policies generally leads to higher expected reward.

3.2 Preliminaries

We consider an infinite-horizon discounted Markov decision process (MDP), defined as $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathcal{P}, r, \rho, \gamma \rangle$, where $\mathcal{S}$ and $\mathcal{A}$ denote the state and action spaces, $\mathcal{P} : \mathcal{S} \times \mathcal{A} \to \Delta_{\mathcal{S}}$ is the transition probability function, $r : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is the reward function, $\rho \in \Delta_{\mathcal{S}}$ is the initial state distribution, and $\gamma \in [0, 1)$ is the discount factor. Throughout this paper, we assume that the state and action spaces are finite but potentially large.

For a fixed $s \in \mathcal{S}$, the policy π induces a distribution $\pi(\cdot|s)$ over the actions. The action-value function $q^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ of policy π is defined as $q^\pi(s, a) := \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) | s_0 = s, a_0 = a]$, with $s_t \sim p(\cdot|s_{t-1}, a_{t-1})$ and $a_t \sim \pi(\cdot|s_t)$. Given an initial state $s \sim \rho$, the corresponding value function is defined as $v^\pi(s) := \mathbb{E}_{a \sim \pi(\cdot|s)}[q^\pi(s, a)]$. The advantage function $\mathfrak{a}^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ induced by π is represented by $\mathfrak{a}^\pi(s, a) := q^\pi(s, a) - v^\pi(s)$. We define $J(\pi) := v^\pi(\rho) = \mathbb{E}_{s \sim \rho}[v^\pi(s)]$ as the expected discounted cumulative reward.

Given the Shannon entropy function $\mathcal{H}(\pi) = -\sum_a \pi(a) \ln(\pi(a))$, if $\tau \geq 0$ is the entropy coefficient, then, the soft (entropy-regularized) counterparts of the above functions [Liu et al., 2024, Vieillard et al., 2020a] are defined as: $v_\tau^\pi(s) := v^\pi(s) + \tau \sum_{t=0}^{\infty} \gamma^t [\mathcal{H}(\pi(\cdot|s_t)) \mid s_0 = s]$, $J_\tau(\pi) := \mathbb{E}_{s \sim \rho}[v_\tau^\pi(s)]$, $q_\tau^\pi(s, a) := \mathbb{E}_{s' \sim \mathcal{P}(\cdot|s,a)} [r(s, a) + \gamma v_\tau^\pi(s')]$ and $\mathfrak{a}_\tau^\pi(s, a) := q_\tau^\pi(s, a) - v_\tau^\pi(s) - \tau \ln(\pi(a|s))$. Note that the soft value and action-value functions, $v_\tau^\pi(s)$ and $q_\tau^\pi(s, a)$, are bounded and lie within the interval $[0, H_\tau]$, where $H_\tau := \frac{1+\tau \ln(A)}{1-\gamma}$ [Liu et al., 2024].

Furthermore, for a fixed $s \in \mathcal{S}$ and an arbitrary pair of policies π_1 and π_2, the *soft* Bellman operator T_τ^π is defined such that: $(T_\tau^{\pi_1} v_\tau^{\pi_2})(s) = \mathbb{E}_{a \sim \pi_1(\cdot|s)} [q_\tau^{\pi_2}(s, a) - \tau \ln(\pi_1(a|s))] = (T^{\pi_1} v_\tau^{\pi_2})(s) + \tau \mathcal{H}(\pi_1(\cdot|s))$ and $(T_\tau^{\pi_1} q_\tau^{\pi_2})(s, a) = (T^{\pi_1} q_\tau^{\pi_2})(s, a) + \tau \mathbb{E}_{s' \sim \mathcal{P}(\cdot|s,a)} \mathcal{H}(\pi_1(\cdot|s'))$. If $\tau = 0$, we refer to the corresponding operator as the *hard* Bellman operator. The objective is to $\max_{\pi \in \Pi} J_\tau(\pi)$, where Π is the set of feasible policies. We denote the optimal entropy-regularized policy by $\pi_\tau^* := \arg\max_\pi J_\tau(\pi)$ and its corresponding value function as v_τ^*.

3.3 General Off-policy Actor-critic Framework

Building on the empirical findings from Section 3.1, we present a general off-policy actor-critic framework that subsumes `DSAC` and yields new actor objectives, including two novel ones. In Section 3.3.1, we focus on the policy evaluation step and instantiate the resulting critic objective. In Section 3.3.2, we focus on two alternative policy updates and instantiate the corresponding actor objectives.

3.3.1 Policy Evaluation

At iteration $t \in [K]$ of the actor-critic algorithm, we evaluate the current policy π_t using the m-step entropy-regularized Bellman operator. The coefficient, $\zeta \geq 0$, controls the strength of

entropy regularization (referred to as the *critic entropy*). The corresponding estimate of the entropy-regularized q function at iteration t is denoted by q_ζ^t and computed recursively as:

$$q_\zeta^0 = q_\zeta^{\pi_0} \quad ; \quad \forall t \geq 1, \quad q_\zeta^t = \mathbb{P}_{[0,H_\tau]}[(T_\zeta^{\pi_t})^m q_\zeta^{t-1}]\,, \tag{3.1}$$

where $\mathbb{P}_{[0,H_\tau]}$ projects each entry onto the $[0, H_\tau]$ interval. As $m \to \infty$, the algorithm exactly evaluates the policy π_t and q_ζ^t converges to the fixed point $q_\zeta^{\pi_t}$. In this special case, setting $\zeta = 0$ recovers the standard q function.

Handling function approximation: Equation (3.1) requires updating the state-action value function for each state and action. In settings where the state or action space is large, this is not computationally feasible and we aim to approximate this update. To this end, we focus on the off-policy setting and define the critic objective using the standard squared loss [Haarnoja et al., 2018a]. Specifically, we use the *replay buffer* $\mathcal{D}_t$ consisting of (state, action, next state, reward) pairs obtained by the policies in the previous iterations. We define ϕ as the parameters of a model (typically a neural network) parameterizing the critic and q_ϕ as the corresponding function. For defining the critic objective, the details of the model are irrelevant and are implicit in the q_ϕ notation. We use Equation (3.1) to construct a *one-step look-ahead target* (corresponding to $m = 1$[1]) and define the critic objective $\mathcal{L}_t(\phi)$ as:

$$\mathbb{E}_{\substack{(s,a,s',r(s,a))\sim\mathcal{D}_t \\ a'\sim\pi_t(\cdot|s')}} \left\| q_\phi(s,a) - \mathbb{P}_{[0,H_\tau]}\left[r(s,a) + \gamma\left[q_\zeta^{t-1}(s',a') - \zeta\,\ln(\pi_t(a'|s'))\right]\right] \right\|_2^2 \tag{3.2}$$

With a slight abuse of notation, by $(s, a, s', r(s, a)) \sim \mathcal{D}_t$, we mean that the tuple is sampled from a discrete distribution over $\mathcal{D}_t$. We set $q_\zeta^t = q_{\phi_t}$ where $\phi_t \approx \arg\min \mathcal{L}_t(\phi)$. Similar to Haarnoja et al. [2018a], in practice, we do not clip the look-ahead target and optimize over the $(s, a, s', r(s, a))$ tuples in a randomly-sampled batch from $\mathcal{D}_t$.

3.3.2 Policy Optimization

At iteration $t \in [K]$ in the actor-critic algorithm, the policy optimization step uses the q function estimates from Section 3.3.1 to update the policy. For a state $s \in \mathcal{S}$, we first compute an intermediate policy $\pi_{t+1/2}$ using two representative methods — (i) the natural policy gradient (`NPG`) or policy mirror descent (PMD) [Kakade, 2001, Xiao, 2022] update and (ii) the recently proposed `SPMA` update [Asad et al., 2025b]. We note that the framework

[1]We focus on $m = 1$ for simplicity. The objective for $m > 1$ can be defined analogously.

is not limited to these choices. With a suitable step-size η_t and appropriate normalization,

$$\pi_{t+1/2}(a|s) \propto \pi_t(a|s)\, \exp(\eta_t\, q^t_\zeta(s,a)) \quad \textbf{(NPG)} \tag{3.3}$$

$$\pi_{t+1/2}(a|s) \propto \pi_t(a|s)\, \left[1 + \eta_t \left(q^t_\zeta(s,a) - v^t_\zeta(s)\right)\right] \textbf{(SPMA)} \tag{3.4}$$

In the special case, when $\zeta = 0$, we note that the SPMA update can use a sufficiently small step-size to avoid an explicit normalization across the actions [Asad et al., 2025b]. On the other hand, the NPG update always requires an explicit normalization across actions to ensure that $\pi_{t+1/2}$ is a valid probability distribution. We now use a proximal update to incorporate entropy-regularization in the policy optimization step. Given $\pi_{t+1/2}$ and the entropy regularization parameter $\tau \geq 0$, the updated policy π_{t+1} can be computed in two alternative ways that aim to find the "closest" policy π to $\pi_{t+1/2}$ while encouraging the resulting policy to have sufficiently high entropy (referred to as the *actor entropy*[2]). Specifically, if $\tau_t := \tau\, \eta_t \geq 0$ is the entropy regularization parameter at iteration t, we use either the forward KL (FKL) or reverse KL (RKL) divergence to measure the proximity between policies. For each state $s \in \mathcal{S}$, the updated policy $\pi_{t+1}(\cdot \mid s)$ is obtained as the solution to one of the following optimization problems:

$$\arg\min_{\pi(\cdot|s)\in\Delta} \mathrm{KL}(\pi_{t+1/2}(\cdot|s)\|\pi(\cdot|s)) - \tau_t \mathcal{H}(\pi(\cdot|s)) \quad \textbf{(FKL)} \tag{3.5a}$$

$$\arg\min_{\pi(\cdot|s)\in\Delta} \mathrm{KL}(\pi(\cdot|s)\|\pi_{t+1/2}(\cdot|s)) - \tau_t \mathcal{H}(\pi(\cdot|s)) \quad \textbf{(RKL)} \tag{3.5b}$$

Note that in the special case when $\tau = 0$, $\pi_{t+1} = \pi_{t+1/2}$. Furthermore, the objective in Equation (3.5b) is convex in π and the resulting update can be obtained in closed form where $\pi_{t+1}(a|s) \propto [\pi_{t+1/2}(a|s)]^{\frac{1}{1+\tau_t}}$ for all (s,a). Combining Equations (3.5a) and (3.5b) with Equations (3.3) and (3.4) gives rise to four possible ways of updating the policy. We instantiate the corresponding actor objectives in the function approximation setting below.

Handling function approximation: Equations (3.5a) and (3.5b) require updating the state-action value function for each state and action. In order to scale to large state-action spaces, we use function approximation in the policy space. Specifically, given the parameters θ of a model parameterizing the actor and $\pi(\theta)$ as the corresponding policy, we define $\Pi_\theta = \{\pi | \exists\, \theta \text{ s.t } \pi = \pi(\theta)\}$ as the set of realizable policies. Similar to Section 3.3.1, the choice of the model is implicit in the $\pi(\theta)$ notation.

Following Haarnoja et al. [2018a], we modify Equations (3.5a) and (3.5b) to optimize (i) only over the states in the replay buffer $\mathcal{D}_t$ and (ii) over the restricted policy class Π_θ, yielding

[2]The actor entropy coefficient τ is independent from the critic entropy coefficient ζ. Intuitively, actor entropy regularizes the distribution over actions at the current state, whereas the critic entropy regularizes the distribution in the next-state.

$\mathbb{E}_{s\sim\mathcal{D}_t}\Big[\mathbb{E}_{a\sim\pi_\theta(\cdot\|s)}\Big[q^t_\zeta(s,a)-\tau\,\ln(\pi_\theta(a\|s))\Big]-\frac{1}{\eta_t}\,\mathrm{KL}(\pi_\theta(\cdot\|s)\,\|\,\pi_t(\cdot\|s))\Big]$	NPG-RKL(τ,ζ)
$\mathbb{E}_{s\sim\mathcal{D}_t}\Big[\mathbb{E}_{a\sim\pi_\theta(\cdot\|s)}\Big[\ln\Big(1+\eta_t(q^t_\zeta(s,a)-v^t_\zeta(s))\Big)-\tau_t\,\ln\pi_\theta(a\|s)\Big]-\mathrm{KL}(\pi_\theta(\cdot\|s)\|\pi_t(\cdot\|s))\Big]$	SPMA-RKL(τ,ζ)
$\mathbb{E}_{s\sim\mathcal{D}_t}\Big[\mathbb{E}_{a\sim\pi_t(\cdot\|s)}\Big[\frac{\exp(\eta_t\,q^t_\zeta(s,a))}{\sum_{a'}\pi_t(a')\,\exp(\eta_t\,q^t_\zeta(s,a'))}\,\ln\Big(\frac{\pi_\theta(a\|s)}{\pi_t(a\|s)}\Big)\Big]+\tau_t\mathcal{H}(\pi_\theta(\cdot\|s))\Big]$	NPG-FKL(τ,ζ)
$\mathbb{E}_{s\sim\mathcal{D}_t}\Big[\mathbb{E}_{a\sim\pi_t(\cdot\|s)}\Big[\frac{1+\eta_t(q^t_\zeta(s,a)-v^t_\zeta(s))}{\sum_{a'}\pi_t(a')(1+\eta_t(q^t_\zeta(s,a')-v^t_\zeta(s)))}\,\ln\Big(\frac{\pi_\theta(a\|s)}{\pi_t(a\|s)}\Big)\Big]+\tau_t\mathcal{H}(\pi_\theta(\cdot\|s))\Big]$	SPMA-FKL(τ,ζ)

Table 3.1: Actor objectives in the proposed framework. (τ,ζ) denotes actor/critic entropy. π_{t+1} for the FKL and RKL variants, respectively: $\arg\min_{\pi\in\Pi_\theta}\sum_{s\sim\mathcal{D}_t}\Big[\mathrm{KL}(\pi_{t+1/2}(\cdot|s)\|\pi(\cdot|s))-\tau_t\mathcal{H}(\pi(\cdot|s))\Big]$,
and $\arg\min_{\pi\in\Pi_\theta}\sum_{s\sim\mathcal{D}_t}\Big[\mathrm{KL}(\pi(\cdot|s)\|\pi_{t+1/2}(\cdot|s))-\tau_t\mathcal{H}(\pi(\cdot|s))\Big]$.

Following Vaswani et al. [2021], Lavington et al. [2023], Tomar et al. [2020], Xiong et al. [2024], we convert the above projection problem into an unconstrained optimization over θ, and form $\ell_t(\theta)$, the corresponding actor objective. We define $\pi_{t+1}=\pi(\theta_{t+1})$ where $\theta_{t+1}\approx\arg\max_\theta\ell_t(\theta)$. For each variant, we append the postfix (τ,ζ) to denote its dependence on the actor and critic entropy. The actor objective $\ell_t(\theta)$ can take one of the four forms, summarized in Table 3.1.

Comparing the objectives: Note that the NPG-RKL and SPMA-RKL objectives differ in the first term, which is linear in the q function for NPG-RKL, while it is logarithmic in the advantage for SPMA-RKL. Similarly, the FKL variants also differ in the first term. Crucially, in the special case of zero critic entropy and $\eta\leq 1-\gamma$, SPMA-FKL$(\tau,0)$ does not require an explicit normalization over the actions [Asad et al., 2025b], making it easier to implement in practice. Finally, we note that for the RKL variants, the expectation is over the actions sampled from π_θ. The objective can be optimized by calculating the full expectation, using importance sampling, or applying the reparameterization trick. On the other hand, the FKL variants involve an expectation over the actions sampled from π_t, simplifying the implementation. In Algorithm 5 in Section B.2, we present the complete actor-critic pseudo-code in the function approximation setting.

Comparison to existing methods: We note that NPG-RKL(τ,τ) recovers the off-policy variant of MDPO studied in Tomar et al. [2020]. In the limit that $\eta_t\to\infty$, NPG-RKL(τ,τ) recovers the original SAC objective in Haarnoja et al. [2018a]. This is intuitive since SAC can be viewed as soft policy iteration. Importantly, NPG-RKL$(\tau,0)$ and $\eta_t\to\infty$ recovers the DSAC variant that demonstrated good empirical performance in Figure 3.1. In the special case of the tabular setting and $\zeta=\tau$ (i.e. the actor and critic entropy is coupled and fixed), the NPG-RKL variant is the same as that proposed in Vieillard et al. [2020a]. On the other hand, the SPMA-RKL(τ,ζ) objective is novel, and has not been studied previously. Moreover,

although NPG-FKL(τ, τ) coincides with the objective of Mei et al. [2019] in the tabular setting, it gives rise to a distinct and novel objective under function approximation.

Finally we note that in the on-policy setting where states are sampled from d^{π_t}, the distribution induced by policy π_t (instead of $\mathcal{D}_t$), setting $\tau = \zeta = 0$ and $m = \infty$ for the critic (corresponding to exact policy evaluation) we can recover the framework in Vaswani et al. [2021], and its instantiations [Tomar et al., 2020, Asad et al., 2025b]. Next, we consider the tabular setting and prove theoretical guarantees for the RKL variants.

3.4 Theoretical Guarantees

We consider the tabular setting and analyze the actor-critic algorithm when using decoupled non-zero critic and actor entropy. For the critic, we consider estimating the q functions using Equation (3.1). For the actor, we consider using the RKL variant in Equation (3.5b) in conjunction with the NPG and SPMA updates in Equations (3.3) and (3.4), and denote the corresponding variants as *soft* NPG and *soft* SPMA respectively.

In Theorem 4 below, we first reduce bounding the sub-optimality in the entropy-regularized value function to a per-state online convex optimization problem. It is important to note that this reduction is independent of the specific actor and critic updates.

Theorem 4 (Generic Reduction with Actor Entropy)**.** *If π^*_τ is the optimal entropy-regularized policy whose value function is v^*_τ, for a q^t_ζ obtained via the policy evaluation scheme at iteration t, define $\epsilon_t := q^t_\zeta - q^{\pi_t}_\tau$. For any sequence of policies $\{\pi_0, \pi_1, \ldots, \pi_{K-1}\}$, if $\bar{\pi}_K$ is the corresponding uniform mixture policy, then,*

$$\left\| v^*_\tau - v^{\bar{\pi}_K}_\tau \right\|_\infty \leq \frac{\|\mathit{Regret}(K)\|_\infty}{K\,(1-\gamma)} + \frac{2\,\sum_{t\in[K]} \|\boldsymbol{\epsilon}_t\|_\infty}{K\,(1-\gamma)}\,,$$

$$\textit{where, } (\mathit{Regret}(K))(s) := \sum_{t=0}^{K-1} \Big[\langle \pi^*_\tau(\cdot|s) - \pi_t(\cdot|s), q^t_\zeta(s,\cdot)\rangle + \tau\big(\mathcal{H}(\pi^*_\tau(\cdot|s)) - \mathcal{H}(\pi_t(\cdot|s))\big)\Big],$$

is the regret incurred on an online optimization problem for each state $s \in \mathcal{S}$.

The above result shows that the sub-optimality of a mixture policy (one that randomly chooses a policy from $\{\pi_0, \pi_1, \ldots, \pi_{K-1}\}$) obtained by using a generic policy optimization method can be bounded in terms of the regret incurred by the method, and the sum of the *policy evaluation errors* incurred in estimating $q^{\pi_t}_\tau$. Since ϵ_t depends on π_t, it depends on the specific policy optimization method, and we now bound it for both soft NPG and soft SPMA.

Corollary 1 (Policy Evaluation Error)**.** *Using the policy evaluation update in Equation* (3.1) *and the soft* NPG *or soft* SPMA *policy update with $\eta_t = \frac{1}{c+\tau\,(t+1)}$ for a constant $c \geq 0$, if*

$\delta(\tau,\zeta) := \frac{|\tau-\zeta|\,\ln(A)}{1-\gamma}$, *the error* ϵ_t *can be bounded for all* $t \in [K]$ *as:*

$$\epsilon_t := \|\boldsymbol{\epsilon}_t\|_\infty = O\bigg(\frac{\gamma^m}{(1-\gamma)^4}\Big(\frac{1}{t}+\frac{1}{K}\Big) + \frac{\delta(\tau,\zeta)}{1-\gamma}\bigg).$$

Note that as m (the number of steps of the Bellman operator) increases, the policy is evaluated more accurately, and ϵ_t decreases. As $m \to \infty$, $\epsilon_t \to O(\delta(\tau,\zeta))$. Moreover, as t increases, ϵ_t decreases and as $t \to K \to \infty$, $\epsilon_t \to O(\delta(\tau,\zeta))$, which quantifies the mismatch between the actor and critic entropy and is equal to zero when $\zeta = \tau$.

Next, we show that both soft NPG and soft SPMA can control the regret for the online optimization problem defined in Theorem 4.

Corollary 2 (Regret Bounds with Actor Entropy)**.** *Suppose* $\pi_0(\cdot|s)$ *is the uniform distribution over actions for state* s*. For a sequence* $\{q_\zeta^t\}_{t=0}^{K-1}$ *with* $\left\|q_\zeta^t\right\|_\infty \le H_\tau$*, the regret for soft* NPG *and soft* SPMA *as defined in Theorem 4, with* $\eta_t = \frac{1}{c+\tau\,(t+1)}$ *for constant* $c \ge 0$ *can be bounded as:*

$$\max_s \bigg|Regret(K)\bigg| = O\bigg(\frac{\ln(K)}{(1-\gamma)^2}\bigg).$$

The actor entropy in Equation (3.5b) makes the online functions strongly-convex and hence both methods incur only a logarithmic regret [Orabona, 2019]. Combining the results in Corollaries 1 and 2 with Theorem 4, we obtain our main theorem for both soft NPG and soft SPMA.

Theorem 5 (Sub-optimality of Soft NPG/Soft SPMA)**.** *Let* π_τ^* *be the optimal entropy-regularized policy with value function* v_τ^**. Consider the soft* NPG *or soft* SPMA *updates with step size* $\eta_t = \frac{1}{c+\tau\,(t+1)}$ *for constant* c *defined in Theorems 13 and 15 and* $\pi_0(\cdot|s)$ *as the uniform policy over actions for all* $s \in \mathcal{S}$*. Using policy evaluation step in Equation* (3.1)*, with* $\delta(\tau,\zeta) := \frac{|\tau-\zeta|\,\ln(A)}{1-\gamma}$*, the resulting uniform mixture policy* $\bar{\pi}_K$ *satisfies the following sub-optimality bound:*

$$\left\|v_\tau^{\bar{\pi}_K} - v_\tau^*\right\|_\infty = \tilde{O}\bigg(\frac{\ln(K)}{K(1-\gamma)^3} + \frac{\gamma^m}{K(1-\gamma)^5} + \frac{\delta(\tau,\zeta)}{(1-\gamma)^2}\bigg).$$

Hence, the sub-optimality can be bounded as $\tilde{O}\left(\frac{1}{K} + \frac{\gamma^m}{K} + \delta(\tau,\zeta)\right)$ up to logarithmic factors. Note that when the actor and critic entropy is coupled ($\zeta = \tau$), both soft NPG and soft SPMA have an $O(1/K)$ convergence for the resulting mixture policy. For the practical variant that uses $m = 1$ and $\zeta = 0$, the above result can achieve an $O(1/K)$ convergence to

an $O(\tau)$ neighbourhood of v_τ^*. All proofs are deferred to Section B.3, where we also explore the case without actor entropy i.e. $\tau = 0$ and prove an $O(1/\sqrt{K})$ rate.

Comparison to existing results: Vieillard et al. [2020a] consider the special case of soft NPG with coupled actor and critic entropy ($\zeta = \tau$) and $m = 1$, and also establish a convergence rate of $O(1/K)$. On the other hand, Xiao [2022, Theorem 4.4] analyze soft NPG with coupled actor and critic entropy ($\zeta = \tau$) and $m = \infty$, and prove an $O(1/K)$ convergence rate. In comparison, our framework can match these convergence rates while being more flexible — we can support $m \in [1, \infty)$ (left as an open question in Vieillard et al. [2020a]), allow for decoupling of actor and critic entropy and support policy optimization methods beyond NPG (e.g., SPMA) that have sublinear regret and Lipschitz policy updates (see the proof of Theorem 11 in Section B.3). In the next section, we empirically study the derived objectives (soft NPG and soft SPMA with decoupled actor and critic entropy, $m = 1$) in the function approximation setting to understand the effect of the proposed design choices.

3.5 Empirical Evaluation

We evaluate DSAC and our proposed framework on various Atari 2600 games [Bellemare et al., 2013]. We first present an ablation study of DSAC, and subsequently investigate the impact of actor and critic entropy and the direction of the KL divergence on the empirical performance of DSAC and the derived objectives. To summarize performance across the 20 Atari games used in the original DSAC and subsequent analyses [Zhou et al., 2022], we follow Agarwal et al. [2021c] and report the IQM of final human-normalized scores over all game-seed pairs, with 95% stratified bootstrap confidence intervals. We further evaluate our hypothesis about DSAC on 20 diverse `stable-retro` games [Poliquin, 2026], which have substantially larger action spaces than Atari in Section B.5.6. For implementation, we follow Tomar et al. [2020], Asad et al. [2025b] and use `stable-baselines3` [Raffin et al., 2021] with five random seeds, reporting per-game average expected returns with 95% confidence intervals across seeds. Experimental details and hyper-parameters for all algorithms are provided in Section B.5.1.

Conclusion 1: Using $\zeta = 0$ Improves DSAC Performance: To isolate the bottleneck contributing to DSAC's poor empirical performance, we fix DQN (a standard baseline for discrete-action environments) as a reference and conduct a targeted ablation of DSAC using a single Q-network. Specifically, we examine the role of critic entropy regularization by comparing $\zeta = 0$ and $\zeta = \tau$, while retaining actor entropy. To set the actor entropy coefficient τ, we use either a fixed grid-searched value or use the adaptive scheme in Christodoulou [2019]. Our results in Figure 3.2 indicate that both $\zeta = 0$ and $\zeta = \tau$ can result in good performance with a *fixed, well-tuned per environment entropy coefficient* τ (red and purple vs. blue). Under this configuration, we also observe that on the discrete classic control tasks Acrobot, MountainCar, and Pendulum reported in Neumann et al. [2018], DSAC performs on

par with GreedyAC, which employs an alternative policy update (see Figures B.1 and B.2 in Section B.5.2). These observations are consistent with entropy-regularized value-based methods [Vieillard et al., 2020b]. In particular, Soft-DQN and M-DQN[3] use a fixed, tuned value of τ.

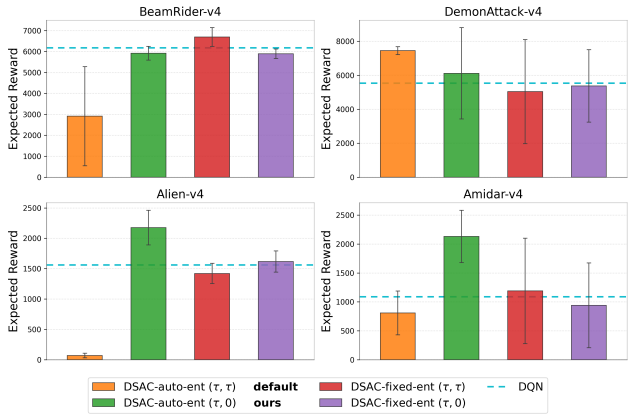


Figure 3.2: **DSAC ablation: sensitivity to critic entropy.** Expected reward after 10M steps. Training DSAC without critic entropy ($\zeta = 0$) and using the default adaptive entropy coefficient loss (auto-ent) from Christodoulou [2019] to set the actor entropy coefficient τ achieves consistently strong performance relative to DQN (green vs. blue). In contrast, the default DSAC ($\zeta = \tau$ with adaptive τ, in orange) performs poorly and improves only with a carefully tuned fixed τ (red).

On the other hand, the adaptive scheme of Christodoulou [2019] sets $\tau_t \approx \arg\min \mathcal{E}_t(\alpha) = \mathbb{E}_{s\sim\mathcal{D}_t}\mathbb{E}_{a\sim\pi_t(.|s)}[-\alpha \ln(\pi_t(a|s)) - \alpha\bar{\mathcal{H}}]$, thereby encouraging the policy entropy to match the fixed *target entropy* $\bar{\mathcal{H}}$. When using this adaptive scheme, $\zeta = 0$ yields performance competitive with DQN (green vs. blue), same as the 20-game results in Figure 3.1, while setting $\zeta = \tau$ (the default setting in DSAC) results in much worse performance (orange vs. blue). Hence, to retain strong empirical performance while avoiding manual tuning, we recommend using the adaptive scheme for actor entropy coefficient τ and setting the critic entropy coefficient $\zeta = 0$.

For the DSAC$(\tau, 0)$ variant with the adaptive scheme of Christodoulou [2019], we test two prior hypotheses for DSAC's poor performance. First, we revisit Zhou et al. [2022], who attribute DSAC's instability partly to coupled actor-critic learning and partly to underestimation from taking the minimum of two Q functions. Their SD-SAC uses a tuned fixed entropy coefficient, an entropy-penalty regularizer limiting policy entropy changes across iterates, and clipped double-average Q-learning. However, Section B.5.3 and Table B.2 shows that taking the minimum of two Q functions does not hurt DSAC$(\tau, 0)$: both single-Q and double-Q variants significantly outperform SD-SAC.

Second, we revisit the claim of Xu et al. [2021] that a fixed target entropy $\bar{\mathcal{H}}$ harms DSAC with adaptive τ, and show that our configuration substantially outperforms their algorithm (Section B.5.4 and Table B.3). We further evaluate our conclusion on hard-exploration Atari

[3]In contrast to DSAC, both Soft-DQN and M-DQN also require tuning an ϵ-greedy parameter and an additional parameter related to entropy regularization [Vieillard et al., 2020b].

Method	IQM	CI Low	CI High
NPG-FKL (0,0), $n = 10$	**1.123**	1.061	1.182
NPG-FKL (0,0), $n = 1$	-0.003	-0.007	0.001
NPG-RKL (0,0), $n = 10$	1.085	0.990	1.172
NPG-RKL (0,0), $n = 1$	-0.006	-0.010	-0.002
SPMA-FKL (0,0), $n = 10$	0.907	0.841	0.983
SPMA-FKL (0,0), $n = 1$	-0.001	-0.005	0.003
SPMA-RKL (0,0), $n = 10$	0.630	0.554	0.709
SPMA-RKL (0,0), $n = 1$	-0.007	-0.011	-0.004
DSAC (0,0), $n = 10$	-0.010	-0.013	-0.007
DSAC (0,0), $n = 1$	-0.006	-0.009	-0.002

Table 3.2: IQM summary of final human-normalized scores with 95% stratified bootstrap confidence intervals in the zero-entropy setting ($\tau = 0$ and $\zeta = 0$): unlike DSAC, objectives from our framework recover strong performance in the absence of entropy regularization when the number of actor gradient updates is increased. Our novel NPG-FKL achieves the strongest performance

games, where DSAC$(\tau, 0)$ outperforms baselines including SD-SAC and TES-SAC (Section B.5.5 and Table B.4), and on stable-retro games with substantially larger action spaces. Our results demonstrate the benefit of disabling critic entropy while using an adaptive actor entropy persists beyond Atari (Section B.5.6).

Conclusion 2: All Objectives from our Proposed Framework Benefit from Setting $\zeta = 0$. From Figure B.7 in Section B.5.7, we observe that when the actor entropy is retained and an adaptive entropy coefficient τ (as in DSAC) is used, all derived objectives with $\zeta = 0$ achieve improved performance relative to their counterparts that use $\zeta = \tau$. *Based on our observations, for the proposed objectives, we recommend using $\zeta = 0$ and adaptive τ for actor entropy.*

Conclusion 3: Entropy Regularization is Only Sometimes Important: We study the effect of disabling entropy regularization for both actor and critic (i.e., $\tau = 0$ and $\zeta = 0$) when the actor objective $\ell_t(\theta)$ is optimized with n gradient steps. For $n = 1$, all objectives, including DSAC, suffer from a sharp drop in policy entropy and degraded performance (see Figures B.9 and B.10), indicating insufficient exploration in the absence of entropy terms. However, unlike DSAC, increasing n alleviates this effect for objectives derived from our proposed framework: setting $n = 10$ preserves higher entropy and yields more stable learning, resulting in improved performance (see Table 3.2). Per-game final expected rewards are shown in Figure B.12 of Section B.5.8.

The gap between DSAC and other objectives can be explained by the lack of regularization in DSAC. In particular, recall that DSAC is a limiting case of NPG-RKL with $\eta_t \to \infty$, and hence does not have the reverse KL regularization term (see the NPG-RKL objective in Table 3.1). Consequently, for $n > 1$, NPG-RKL$(0, 0)$ benefits from additional optimization steps under KL regularization, whereas DSAC$(0, 0)$ lacks both the KL and entropy terms.

Conclusion 4: Objectives from Our Proposed Framework are Comparable to Baselines: To evaluate the empirical effectiveness of the derived objectives, we compare against established baselines PPO [Schulman et al., 2017b], DQN [Mnih et al., 2013], and the

Method	IQM	CI Low	CI High
NPG-FKL $(\tau, 0)$	**1.2193**	1.1603	1.2827
NPG-RKL $(\tau, 0)$	1.0092	0.8784	1.1430
SPMA-FKL $(\tau, 0)$	1.0088	0.8970	1.1301
SPMA-RKL $(\tau, 0)$	1.0529	0.9659	1.1561
DSAC $(\tau, 0)$	0.9226	0.8228	1.0288
Adam LMCDQN (Baseline)	1.0363	0.9423	1.1265
DQN (Baseline)	0.7827	0.7563	0.8091
PPO (Baseline)	0.6835	0.6258	0.7311

Table 3.3: IQM summary of final human-normalized scores with 95% stratified bootstrap confidence intervals: comparing our objectives with actor entropy enabled ($\tau \neq 0$), critic entropy disabled ($\zeta = 0$), and $n = 1$ actor update against baselines. Our novel `NPG-FKL`$(\tau, 0)$ performs best.

recent value-based method `Adam LMCDQN` [Ishfaq et al., 2023]. We select `Adam LMCDQN` as a baseline for three reasons. First, it matches or outperforms strong value-based baselines, including `NoisyNet DQN` [Fortunato et al., 2017], `C51` [Bellemare et al., 2017], `QRDQN` [Dabney et al., 2017], and `IQN` [Dabney et al., 2018], across a broad set of Atari games. Second, `Adam LMCDQN` avoids ϵ-greedy exploration, which is known to be brittle to tune [Hessel et al., 2018]. Third, the method is implemented as a modification of Adam [Kingma and Ba, 2015], making it straightforward to integrate into the same `stable-baselines3` codebase used for all methods in our framework. This choice allows for a controlled comparison that isolates algorithmic differences rather than implementation artifacts.

The IQM results across all games in Table 3.3 show that our novel `NPG-FKL`$(\tau, 0)$ achieves the strongest overall performance. Per-game final expected rewards are provided in Section B.5.9. Full training curves are shown in Figures B.14–B.16 in Appendix B.5.10. We emphasize that our goal is to evaluate the objectives derived from our principled framework, rather than to construct a fully compositional agent such as `Rainbow` [Hessel et al., 2018]. Standard enhancements, including prioritized replay, multi-step returns, distributional critics, dueling architectures, NoisyNets, and double Q-learning might further improve the performance. However, such modifications are orthogonal to our contributions, and we leave them to future work.

Conclusion 5: Empirical Results Tend to Favour the Forward KL Direction: Chan et al. [2022] have studied the impact of the KL direction in the discrete-action setting when the intermediate policy is Boltzmann i.e. $\pi_{t+1/2} \propto \exp(q^t_\tau(s, a)/\tau)$, and have reported no significant differences in the performance. We revisit this question in the context of our actor objectives, evaluating `SPMA` and `NPG` under two regimes: (i) $\zeta = 0, \tau = 0$ and (ii) $\zeta = 0$, $\tau \neq 0$ (i.e., only actor entropy). For (i), we use $n = 10$ (as $n = 1$ performs poorly) and observe that using the forward KL yields a clear advantage for `SPMA`, while the difference is small for `NPG` (Table 3.2). For (ii), forward and reverse KL perform similarly for `SPMA`, whereas `NPG` benefits more from the forward KL direction (Table 3.3).

3.6 Conclusion and Future Work

We revisited off-policy actor-critic design in discrete-action settings, where DSAC performs poorly. By decoupling actor and critic entropy, we identified a DSAC variant that substantially improves over the default. Motivated by this, we introduced a flexible off-policy actor-critic framework that subsumes DSAC and yields novel actor objectives, including SPMA-RKL and NPG-FKL, that remain robust even without entropy regularization or explicit exploration.

Future work includes: (i) analyzing these actor objectives with function approximation to guide algorithm design; (ii) developing adaptive schemes for the actor step size η_t; (iii) further studying the objectives without entropy regularization; (iv) incorporating enhancements from compositional algorithms such as Rainbow; and (v) extending the framework to continuous-action settings.

Towards optimistic actor-critic methods with parametric policies. Beyond the design of actor and critic updates, learning also depends on which experience the agent collects. The tabular analysis in this chapter uses policy-evaluation updates over all state-action pairs, leaving open how interaction provides the information required for these updates. Entropy regularization promotes action diversity, but alone need not ensure efficient exploration [Cesa-Bianchi et al., 2017, O'Donoghue, 2023]. Recall the N-chain in Figure 1.1: encouraging diverse actions within the visited prefix need not direct the agent toward the uncertain continuation and its distant reward. Optimism uses uncertainty in estimated values to favour actions with high plausible returns, so that the possibility of a valuable continuation can guide the actor's exploration. Incorporating this principle into a practical actor-critic method raises a further challenge: existing optimistic actor-critic methods often use implicit policies that can depend on stored past critics [Liu et al., 2023, Cassel and Rosenberg, 2024, Tan et al., 2025]. Extending these methods to explicitly parameterized actors remains an open problem, even in the simple setting of finite-horizon linear MDPs. The next chapter therefore studies optimistic actor-critic methods with explicit parametric policies in finite-horizon linear MDPs. It combines a parametric actor with Langevin Monte Carlo critic updates to obtain sample-efficiency guarantees that account for exploration in both the on-policy and off-policy settings. Furthermore, the chapter develops a preliminary adaptation of the proposed optimistic actor-critic algorithm for large-scale Atari environments.

Chapter 4

Optimistic Actor-Critic with Parametric Policies

4.1 Introduction

Reinforcement learning (RL) is a general framework for sequential decision making under uncertainty and has been successful in various real-world applications, such as robotics [Kober et al., 2013] and large language models [Uc-Cetina et al., 2023]. Policy Gradient (PG) methods [Williams, 1992, Sutton et al., 1999, Kakade, 2001, Schulman et al., 2017a] are an important class of algorithms that assume a differentiable parameterization of the policy, and directly optimize the policy parameters using the return from interacting with the environment. PG methods are widely used in practice as they can easily handle function approximation or structured state-action spaces. However, since the environment is typically stochastic in practice, the estimated returns usually have high variance, resulting in poor sample efficiency [Dulac-Arnold et al., 2019].

Actor-critic (AC) methods [Konda and Tsitsiklis, 1999, Peters et al., 2005, Bhatnagar et al., 2009] alleviate this issue by using value-based approaches in conjunction with PG methods. In particular, they utilize a critic that estimates the policy's value and an actor that performs PG to improve the policy towards obtaining higher returns. These AC methods have been proven to be empirically successful in both on-policy [Schulman et al., 2015a, 2017b] and off-policy [Lillicrap et al., 2015, Fujimoto et al., 2018, Haarnoja et al., 2018a] settings.

Subsequently, there have been many attempts to provide a theoretical understanding of actor-critic methods, especially in the presence of function approximation [Cai et al., 2020, Zhong and Zhang, 2023, Liu et al., 2023]. However, there are two prevalent issues that result in mismatches between theory and practice: the studied methods either (i) do not consider strategic exploration in a systematic manner or (ii) analyze complicated and impractical

Algorithm	Sample Complexity (On-Policy)	Sample Complexity (Off-Policy)	Policy Param.	Clipping Q-Function	Computational Cost for Policy Inference
Liu et al. [2023]	$\widetilde{\mathcal{O}}(\frac{1}{\epsilon^4})$	✗	implicit	yes	$\mathcal{O}(d_{\boldsymbol{c}}^2\, H\, \lvert\mathcal{A}\rvert\, \epsilon^{-2})$
Sherman et al. [2023] Cassel and Rosenberg [2024]	✗	$\widetilde{\mathcal{O}}(\frac{1}{\epsilon^2})$	implicit	no	$\mathcal{O}(d_{\boldsymbol{c}}^2\, H\, \lvert\mathcal{A}\rvert\, \epsilon^{-2})$
Ours	$\widetilde{\mathcal{O}}(\frac{1}{\epsilon^4})$	$\widetilde{\mathcal{O}}(\frac{1}{\epsilon^2})$	explicit	yes	$\mathcal{O}(d_{\boldsymbol{a}}^2\, H\, \lvert\mathcal{A}\rvert)$

Table 4.1: Comparison with the state-of-the-art algorithms for episodic finite-horizon linear MDPs (with feature dimension $d_{\boldsymbol{c}}$ and horizon H). The sample complexity refers to the number of interactions required for outputting an ϵ-optimal policy, whereas the cost of policy inference refers to the per-episode cost for interacting with the environment for H steps. Our proposed algorithm matches the optimal sample efficiency in both the on- and off-policy settings. Furthermore, in contrast to existing works, our algorithm employs an explicit policy parameterization (log-linear policies with dimension $d_{\boldsymbol{a}}$), resulting in lower computational cost for policy inference.

variants of the algorithm. In particular, much of the literature makes unrealistic assumptions to avoid dealing with exploration, a central challenge in RL. For instance, existing works on PG methods [Agarwal et al., 2021a, Yuan et al., 2023, Alfano and Rebeschini, 2022b, Asad et al., 2025b] obtain convergence rates that involve a mismatch ratio between the optimal policy and the initial state distribution. These results are only meaningful if the mismatch ratio is bounded. However, a bounded mismatch ratio indicates that the initial state distribution already provides a good coverage over the state space, thereby sidestepping the exploration problem. Within actor-critic methods, some early analyses make assumptions on the reachability of the state-action space or the coverage of collected data [Abbasi-Yadkori et al., 2019, Neu et al., 2017, Bhandari and Russo, 2024, Agarwal et al., 2021a, Cen et al., 2022, Gaur et al., 2024], which again imply that the state-action space is already relatively easy to explore. Follow-up works [Hong et al., 2023, Fu et al., 2021, Xu et al., 2020, Cayci et al., 2024] assume a bounded mismatch ratio, while others [Khodadadian et al., 2022, Gaur et al., 2023] require mixing assumptions on the induced Markov chain.

On the other hand, recent works [Cai et al., 2020, Jin et al., 2021, Zanette et al., 2021, Zhong and Zhang, 2023, Agarwal et al., 2023, He et al., 2023, Liu et al., 2023, Sherman et al., 2023, Cassel and Rosenberg, 2024, Tan et al., 2025] tackle the exploration issue directly. However, the algorithms analyzed are significantly different from those implemented in practice. Much of this body of work studies AC methods that use the natural policy gradient (`NPG`) update for policy optimization. However, the canonical implementation of the `NPG` update does not consider an explicit policy parameterization. Instead, the update involves constructing "implicit" policies *on the fly* using all previously stored Q-functions. This makes it computationally expensive to sample from these policies and use them to interact with the environment. In contrast, the algorithms used in practice typically employ

explicitly parameterized complex models as learnable policies and optimize them with gradient descent-based methods. Therefore, we aim to address the following question:

Can we design a provably efficient actor–critic algorithm with parametric policies for linear MDPs in both on- and off-policy settings?

Contributions. We answer the above question affirmatively and make the following contributions.

1. General framework with an explicitly parameterized actor. In Section 4.3, we propose a general optimistic actor-critic framework that utilizes an explicitly parameterized policy. We apply this framework in the setting of linear function approximation for both the environment (i.e., linear MDP [Jin et al., 2020]) and the policy (i.e., log-linear policy class). In Section 4.4, we propose an actor algorithm that learns a log-linear policy by solving a specific regression problem at each iteration. This allows us to directly control the error between the explicitly parametrized policy and the implicit policy induced by `NPG`. Using this error bound in conjunction with the well-established theoretical results of `NPG` [Hazan et al., 2016, Szepesvári, 2022] enables us to analyze the performance of the parameterized actor. We show that the proposed algorithm benefits from a substantially improved memory complexity, while retaining similar theoretical guarantees.

2. LMC critic for practical strategic exploration. In Section 4.5, instead of constructing `UCB` bonuses [Jin et al., 2020], which are ubiquitous within prior works [Zhong and Zhang, 2023, Liu et al., 2023, Sherman et al., 2023, Cassel and Rosenberg, 2024], we adopt a more practical approach. We employ Langevin Monte Carlo (`LMC`) [Welling and Teh, 2011] to update the critic parameters at each episode. Unlike `UCB`-based approaches that require computing confidence sets at every episode, `LMC` simply perturbs (by Gaussian noise) the gradient descent update on the critic loss. This gradient descent-based approach is both easier to implement [Ishfaq et al., 2025] and to extend to general function approximation [Ishfaq et al., 2024b]. Furthermore, the `LMC` algorithm directly leads to an optimistic estimate of the Q-function that has similar guarantees as `UCB` bonuses. Nevertheless, previous works have only successfully designed provably efficient algorithms for solving multi-armed bandits [Mazumdar et al., 2020], contextual bandits [Xu et al., 2022], and linear MDPs via value-based methods [Ishfaq et al., 2024a]. Our paper is the first to analyze an `LMC`-based approach in the context of policy optimization.

3. End-to-end theoretical guarantees for actor-critic. In Section 4.6, we analyze the proposed actor-critic framework in both the on-policy and off-policy settings without making any assumptions on the mismatch ratio or data coverage. In particular, in the on-policy

setting, we prove that our method requires $\widetilde{\mathcal{O}}(\epsilon^{-4})$ samples to learn an ϵ-optimal policy. This matches the result in [Liu et al., 2023] that uses an implicit NPG policy in conjunction with UCB bonuses. On the other hand, we also prove that our framework can attain a sample complexity of $\widetilde{\mathcal{O}}(\epsilon^{-2})$ in the off-policy setting. This matches the results of Sherman et al. [2023], Cassel and Rosenberg [2024], but with a far less complicated algorithm design.

We thus demonstrate that our proposed algorithm is both sample-efficient and aligned with practice.

4.2 Preliminaries

In this section, we introduce the finite-horizon linear MDP setting and the log-linear policy class.

Finite-Horizon Linear MDP. A finite-horizon MDP is a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathbb{P}, r, H)$ where $\mathcal{S}$ denotes the state space, $\mathcal{A}$ is the action set, and $H \in \mathbb{Z}_+$ is the length of the horizon. $\mathbb{P} = \{\mathbb{P}_h\}_{h\in[H]}$ is a set of time-dependent transition kernels, and $r = \{r_h\}_{h\in[H]}$ denotes a sequence of reward functions. We assume that the state space $\mathcal{S}$ is a (possibly infinite) measurable space, whereas $\mathcal{A}$ is a finite set with cardinality $|\mathcal{A}|$. $\mathbb{P}_h(\cdot \mid s, a) \in \Delta(\mathcal{S})$ is the distribution over states when taking action $a \in \mathcal{A}$ in state $s \in \mathcal{S}$ at step $h \in [H]$, and $r_h(s, a) \in [0, 1]$ is the corresponding reward. Additionally, for any given function $V : \mathcal{S} \to \mathbb{R}$, we define that $[\mathbb{P}_h V_{h+1}](s, a) \coloneqq \mathbb{E}_{s'\sim\mathbb{P}_h(\cdot|s,a)} V_{h+1}(s')$.

The agent interacts with the environment by starting at an initial state (w.l.o.g., fixed to be $s_1 \in \mathcal{S}$). At step h, the agent first observes the current state $s_h \in \mathcal{S}$, then takes an action $a_h \in \mathcal{A}$ and receives the reward $r_h(s_h, a_h)$. After that, the agent transitions to $s_{h+1} \sim \mathbb{P}_h(\cdot \mid s_h, a_h)$. The agent follows a given policy $\pi : [H] \times \mathcal{S} \mapsto \Delta(\mathcal{A})$ in which $\pi_h(\cdot \mid s) \in \Delta(\mathcal{A})$ is the probability distribution over $\mathcal{A}$ in state s at step h. To quantify the performance of any policy π, we define the value function as $V_h^\pi(s) \coloneqq \mathbb{E}_{\pi,\mathbb{P}}[\sum_{\tau=h}^{H} r_\tau(s_\tau, a_\tau) \mid s_h = s]$, and the corresponding state-action value function is defined as $Q_h^\pi(s, a) \coloneqq \mathbb{E}_{\pi,\mathbb{P}}[\sum_{\tau=h}^{H} r_\tau(s_\tau, a_\tau) \mid s_h = s, a_h = a]$, where the expectation is with respect to the randomness in the stochastic policy and the transition dynamics. The value function (resp. Q-function) corresponds to the expected cumulative rewards when starting in state s (resp. state-action (s, a)) at step h, and subsequently following the policy π until reaching step H.

We assume that both $\mathbb{P}$ and r are unknown to the agent. In order to efficiently learn these quantities, we consider the linear MDP assumption [Jin et al., 2020] where both the transition kernel and the reward function are assumed to be linear functions of given features.

Definition 1 (Linear MDP)**.** *A finite-horizon MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathbb{P}, r, H)$ is a linear MDP with a feature map $\phi : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}^{d_c}$ if the following holds. There exist H signed measures $\psi_h : \mathcal{S} \to \mathbb{R}^{d_c}$ and $\upsilon_h : \mathbb{R}^{d_c}$ such that $\mathbb{P}_h(s' \mid s, a) = \langle \phi(s, a), \psi_h(s') \rangle$ and $r_h(s, a) = \langle \phi(s, a), \upsilon_h \rangle$*

for all h, s, and a. It should also satisfy the following constraints: $\|\phi(s,a)\|_2 \leq 1$ and $\|\upsilon_h\|_2 \leq \sqrt{d_{\boldsymbol{c}}}$ for all h, s, and a. Additionally, for any measurable function $V : \mathcal{S} \to [0,1]$, $\left\|\int_{s\in\mathcal{S}} V(s)\,\psi_h(s)\,\mathrm{d}s\right\|_2 \leq \sqrt{d_{\boldsymbol{c}}}$.

According to Jin et al. [2020, Proposition 2.3], for a linear MDP and any policy π, Q_h^π is a linear function of the features: for all (h,s,a), there exists a $w_h \in \mathbb{R}^{d_{\boldsymbol{c}}}$ such that $Q_h^\pi(s,a) = \langle \phi(s,a), w_h \rangle$.

Learning Objective. For this linear MDP setting, we assume that only ϕ is given to the learner whereas ψ and υ are not. The agent sequentially interacts with the environment for T episodes and aims to minimize the *cumulative regret* defined as $\mathtt{Reg}(T) \coloneqq \sum_{t=1}^{T} [V_1^\star(s_1) - V_1^{\pi_t}(s_1)]$, where $V_1^\star \coloneqq V_1^{\pi^\star} \coloneqq \sup_\pi V_1^\pi$ is the value function of the optimal policy $\pi^\star \coloneqq \arg\sup_\pi V_1^\pi(s_1)$. Equivalently, if $\overline{\pi}^T$ denotes the mixture policy that picks a policy among $\{\pi^1, \ldots, \pi^T\}$ uniformly randomly, we aim to learn an ϵ-optimal $\overline{\pi}^T$, i.e., its *optimality gap* ($\mathtt{OG}$) is bounded such that

$$\mathtt{OG}(T) \coloneqq \mathbb{E}\left[V_1^\star(s_1) - V_1^{\overline{\pi}^T}(s_1)\right] = \frac{\mathtt{Reg}(T)}{T} \leq \widetilde{\mathcal{O}}(\epsilon)\,, \tag{4.1}$$

where the expectation is taken with respect to the randomness of the mixture policy.

Log-Linear Policy. We consider a restricted policy class Π_{lin} consisting of *log-linear policies*. Log-linear policies are represented using the softmax function with linear function approximation. In particular, a log-linear policy is defined as follows: for all $h \in [H]$,

$$\pi_h(a \mid s, \theta) = \frac{\exp(z_h(s,a \mid \theta_h))}{\sum_{a'\in\mathcal{A}} \exp(z_h(s,a' \mid \theta_h))}\,, \tag{4.2}$$

where $z_h(s,a \mid \theta_h) = \langle \varphi(s,a), \theta_h \rangle$ represents the logits parameterized by θ_h, and $\varphi : \mathcal{S}\times\mathcal{A} \mapsto \mathbb{R}^{d_{\boldsymbol{a}}}$ are policy features given to the learner. W.l.o.g, we assume that $\|\varphi(s,a)\| \leq 1$ for all s and a. For convenience, we use the shorthand $\pi(\theta) : [H]\times\mathcal{S} \to \Delta(\mathcal{A})$ to refer to the log-linear policy corresponding to the parameters θ.

4.3 A General Actor-Critic Framework with Parametric Policies

In this section, we start by introducing our general optimistic actor-critic framework as shown in Algorithm 2. Starting with a uniform policy π^1, at the beginning of every learning episode $t \in [T]$, the agent interacts with the environment using policy π^t (Line 4). Our framework allows for collecting data from the environment in either an on-policy or off-policy fashion. In the on-policy setting, at episode t, the agent collects N fresh trajectories $\mathcal{D}^t$ by interacting with the environment using the current policy π^t. On the other hand, in the

Algorithm 2: Actor-Critic with Parametric Policies

1: **Input**: number of update steps T, data collection batch size N (only for on-policy)
2: set $\mathcal{D}^0 \leftarrow \emptyset$, $w_h^1 \leftarrow \mathbf{0}$, $\pi_h^1(\cdot \mid s) \leftarrow \text{Unif}(\mathcal{A}) \quad \forall (h, s)$
3: **for** $t = 1, \ldots, T-1$ **do**
4: $\quad \mathcal{D}^t \leftarrow \begin{cases} \text{On-Policy:} & \left\{N \text{ fresh traj. } \overset{\text{i.i.d.}}{\sim} \pi^t\right\} \\ \text{Off-Policy:} & \mathcal{D}^{t-1} \cup \left\{1 \text{ traj. } \sim \pi^t\right\} \end{cases}$
5: $\quad w^{t+1} \leftarrow \texttt{CRITIC_UPDATE}(\mathcal{D}^t, \pi^t, w^t)$
6: $\quad \theta^{t+1} \leftarrow \texttt{ACTOR_UPDATE}(w^{t+1}, \theta^t)$
7: $\quad \pi^{t+1} = \pi(\theta^{t+1})$
8: **end for**
9: **Return**: mixture policy $\overline{\pi}^T$

off-policy setting, at episode t, the agent collects only 1 trajectory from the environment using π^t. However, the agent stores all the historical data collected by the previous policies, and hence, $\mathcal{D}^t$ consists of t trajectories, each collected by $\pi^1, \ldots, \pi^t$ respectively.

The *critic* uses the collected data and estimates an (optimistic) Q-function via learning the critic parameters $w^{t+1} \in [H] \times \mathbb{R}^{d_c}$ (Line 5). The *actor* then uses the estimated Q-function, and updates the parameters of $\theta^t \in [H] \times \mathbb{R}^{d_a}$ of the log-linear policy (Line 6). The updated log-linear policy is denoted by π^{t+1} (Line 7), and is used to collect data in the next episode.

Given this general framework, we will next instantiate the actor in Section 4.4 and the critic in Section 4.5.

4.4 Instantiating the Actor: Projected Natural Policy Gradient

In this section, we instantiate the actor using natural policy gradient (`NPG`) with parametric policies and analyze its behavior. In particular, in Section 4.4.1, we devise an algorithm that projects the standard `NPG` update onto the class of realizable policies. In Section 4.4.2, we analyze and control the errors induced by the projection step. Finally, in Section 4.4.3, we put everything together and instantiate the complete actor algorithm for the log-linear policy class.

4.4.1 Projected Natural Policy Gradient

At episode $t \in [T]$, given $\widehat{Q}_h^t$, the estimated Q-function, `NPG` updates the policy as: for every $(h, s) \in [H] \times \mathcal{S}$,

$$\pi_h^{t+1}(\cdot|s) \propto \pi_h^t(\cdot|s) \, \exp(\eta \, \widehat{Q}_h^t(s, \cdot)) \tag{4.3}$$

with the corresponding normalization across $\mathcal{A}$. Existing work on policy optimization in linear MDPs [Liu et al., 2023, Sherman et al., 2023, Cassel and Rosenberg, 2024] uses NPG to update the actor because of its favorable theoretical properties. Importantly, *these results do not consider any explicit parameterization for the actor*. Directly implementing the update in Equation (4.3) requires $\mathcal{O}(|\mathcal{S}||\mathcal{A}|)$ memory, and is therefore impractical with large state-action spaces.

Consequently, existing work uses the following equivalent form of the NPG update: for every $(h, s) \in [H] \times \mathcal{S}$,

$$\pi_h^{t+1}(\cdot|s) \propto \pi_h^1(\cdot|s) \exp\left(\eta \sum_{i=1}^t \widehat{Q}_h^i(s, \cdot)\right), \tag{4.4}$$

which characterizes the policy implicitly. In particular, at episode t, for any (h, s, a), we can compute the policy *on the fly* if we have access to the sum of all the parameterized Q-functions up to episode t. However, in existing works [Liu et al., 2023, Sherman et al., 2023, Cassel and Rosenberg, 2024], the sum of parameterized Q functions cannot be stored in a succinct manner, and they require storing *all* the parameterized Q functions. Consequently, when interacting with the environment using such an implicit policy, these works suffer from an extensive per-episode computational cost of $\mathcal{O}(d_{\mathcal{C}}^2 H |\mathcal{A}| T)$. Therefore, the resulting algorithm is far from practice that typically uses an explicit (and often sophisticated) actor parameterization.

To alleviate these issues, we aim to compute a policy that is (i) realizable by the explicit actor parameterization and (ii) provably approximates the policy induced by the NPG update in Equation (4.3) (referred to as the *implicit policy*). To this end, we use a projected NPG update:

$$\pi_h^{t+1}(\cdot \mid s) = \mathrm{Proj}_{\Pi}\left[\frac{\pi_h^t(\cdot \mid s) \exp(\eta \widehat{Q}_h^t(s, \cdot))}{\sum_{a'} \pi_h^t(a' \mid s) \exp(\eta \widehat{Q}_h^t(s, a'))}\right], \tag{4.5}$$

where Proj is the projection operator, which will be instantiated subsequently in Section 4.4.2.

When theoretically analyzing policy optimization methods, an important intermediate result is the bound on the regret for a specific online linear optimization problem. For the standard NPG update in Equation (4.3), this regret can be bounded by $\widetilde{\mathcal{O}}(\sqrt{T})$ [Hazan et al., 2016, Szepesvári, 2022]. In the following Theorem, we analyze the effect of the projection operator and bound the regret for the projected NPG.

Theorem 6. *Given a sequence of linear functions $\{\langle p^t, g^t\rangle\}_{t\in[T]}$ for a sequence of vectors $\{g^t\}_{t\in[T]}$ where for any $t \in [T]$, $p^t \in \Delta(\mathcal{A})$, $g^t \in \mathbb{R}^{|\mathcal{A}|}$, and $\|g^t\|_\infty \le H$. Consider $p^{t\in[T]}$ where p^1 is the uniform distribution, and for all $t \in [T]$,*

$$p^{t+1/2} = \arg\min_{p\in\Delta_A} \left\{\left\langle p, -\eta\, g^t\right\rangle + \mathrm{KL}(p \,\|\, p^t)\right\}, \tag{4.6}$$

$$p^{t+1} = \mathrm{Proj}_{\Pi}(p^{t+1/2})\,. \tag{4.7}$$

Let $\epsilon^t \coloneqq \mathrm{KL}(u \,\|\, p^{t+1}) - \mathrm{KL}(u \,\|\, p^{t+1/2})$ *be the projection error induced by Equation* (4.7)*. Then, for any comparator* $u \in \Delta(\mathcal{A})$*, it holds that*

$$\sum_{t=1}^{T} \left\langle u - p^t, g^t \right\rangle \leq \frac{\log|\mathcal{A}| + \sum_{t=1}^{T} \epsilon^t}{\eta} + \frac{\eta\, H^2\, T}{2}. \tag{4.8}$$

The update in Equation (4.6) with $p^t = \pi_h^t(\cdot|s)$ is equivalent to the standard NPG update in Equation (4.3) [Xiao, 2022]. Using this lemma for each state s and step h, with $g^t = \widehat{Q}_h^t(s,\cdot)$ and an appropriate choice of η gives the following regret bound[1] for the projected NPG:

$$\textstyle\sum_{t=1}^{T} \sum_{h=1}^{H} \max_{s\in\mathcal{S}} \left\langle \pi_h^\star(\cdot|s) - \pi_h^t(\cdot|s), \widehat{Q}_h^t(s,\cdot) \right\rangle \tag{4.9}$$

$$\leq \mathcal{O}\left(H^2 \sqrt{\log|\mathcal{A}|}\,\sqrt{T} + H^2\,\sqrt{\bar{\epsilon}}\,T\right), \tag{4.10}$$

where $\bar{\epsilon} \coloneqq \max_{t,s,h} \epsilon_h^t(s)$ is the largest error across all t, s, and h. For the NPG in Equation (4.3) without projection, $\bar{\epsilon} = 0$ and the above result recovers the standard regret bound for NPG. The above lemma suggests that by choosing the projection operator carefully and controlling the projection errors, we can bound the regret.

4.4.2 Controlling the Projection Error for Log-Linear Policies

To bound the projection error in Theorem 6, one could choose that $\mathrm{Proj}_{\Pi}(p) = \arg\min_p \mathrm{KL}(u \,\|\, p) - \mathrm{KL}(u \,\|\, p^{t+1/2})$, and hence directly control ϵ_h^t. However, this results in a non-convex optimization problem. Consequently, we instead choose Proj to minimize the following regression loss in the logit space: $\frac{1}{2}\left\|z - (z^t + \eta\, g^t)\right\|$ where z^t is the logit corresponding to p^t such that $p^t \propto \exp(z^t)$. For the projected NPG with log-linear policies, we aim to minimize the sum of such regression losses (across all $(s,a) \in \mathcal{S}\times\mathcal{A}$) at episode t and step h, and obtain the loss function:

$$\ell_h^t(\theta) = \frac{1}{2} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \left[\left\langle \varphi(s,a), \theta - \widehat{\theta}_h^t \right\rangle - \eta\, \widehat{Q}_h^t(s,a)\right]^2, \tag{4.11}$$

where $\widehat{Q}_h^t(s,a)$ is the estimated Q-function from the critic. As a regression problem, this actor loss can be easily optimized via gradient descent-based methods.

However, note that the above actor loss requires a minimization over the entire state-action space, which may be impractical. Therefore, we propose to construct a good and

[1]This generalized regret bound holds for any other mirror descent-based policy optimization method (e.g., SPMA [Asad et al., 2025b] in Section C.2.3), but we discuss NPG within the main text for the ease of exposition.

preferably small subset $\mathcal{D}_{\text{exp}} \subset \mathcal{S} \times \mathcal{A}$ along with a corresponding distribution $\rho_{\text{exp}} \in \Delta(\mathcal{D}_{\text{exp}})$ that offers good coverage of the feature space. Given $\mathcal{D}_{\text{exp}}$ and ρ_{exp} (the construction of which will be detailed subsequently), we instantiate the actor loss, $\widetilde{\ell}_h^t(\theta)$, in terms of a *logit-matching* regression:

$$\widetilde{\ell}_h^t(\theta) = \frac{1}{2} \sum_{(s,a) \in \mathcal{D}_{\text{exp}}} \rho_{\text{exp}}(s,a) \left[\langle \varphi(s,a), \theta \rangle - \widehat{Z}_h^t(s,a) \right]^2 , \tag{4.12}$$

$$\text{where } \widehat{Z}_h^t(s,a) \coloneqq \left\langle \varphi(s,a), \widehat{\theta}_h^t \right\rangle + \eta \, \widehat{Q}_h^t(s,a) \, . \tag{4.13}$$

In order to show that optimizing the above actor loss can indeed bound the projection error, we require the following assumptions. We assume that the given policy features φ are expressive enough to control the bias when minimizing $\ell_h^t(\theta)$.

Assumption 4 (Bias)**.** $\min_\theta \sup_{t,h} \widetilde{\ell}_h^t(\theta) \le \epsilon_{\text{bias}}$

In practice, ϵ_{bias} can be controlled by choosing high-dimensional features (e.g., $d_{\boldsymbol{a}} \gg d_{\boldsymbol{c}}$) or a sufficiently expressive policy class (e.g., neural network).

Next, we assume the loss $\widetilde{\ell}_h^t(\theta)$ is sufficiently minimized.

Assumption 5 (Optimization Error)**.** *Suppose θ_h^t is obtained by minimizing $\widetilde{\ell}_h^t(\theta)$ in the critic update.* $\sup_{t,h} \left| \widetilde{\ell}_h^t(\theta_h^t) - \min_\theta \widetilde{\ell}_h^t(\theta) \right| \le \epsilon_{\text{opt}}$.

In practice, minimizing $\widetilde{\ell}_h^t(\theta)$ by K_t steps of gradient descent ensures that $\epsilon_{\text{opt}} \le \mathcal{O}(\exp(-K_t))$. Given these two mild assumptions, we can then proceed to bound the projection error. Using Theorem 6 for the projected `NPG` update at state s, step h, and setting $u = \pi^\star(\cdot \mid s)$, the projection error $\epsilon_h^t(s)$ can be bounded as follows.

Lemma 1. *Let $\overline{\varphi}_G \coloneqq \sup_{(s,a) \in \mathcal{S} \times \mathcal{A}} \|\varphi(s,a)\|_{G^{-1}}$ where $G \coloneqq \sum_{(s,a) \in \mathcal{D}_{\text{exp}}} \rho_{\text{exp}}(s,a) \, \varphi(s,a) \, \varphi(s,a)^\top$. Under Assumptions 4 and 5, it holds that*

$$\left| \epsilon_h^t(s) \right| \le \bar{\epsilon} \coloneqq 2 \left(\overline{\varphi}_G + 1 \right) \sqrt{\epsilon_{\text{bias}}} + 2 \sqrt{\epsilon_{\text{opt}}} \quad \forall (t,h,s) \, . \tag{4.14}$$

The above lemma is true for any choice of $\mathcal{D}_{\text{exp}}$ and ρ_{exp}, and suggests that if we can control $\overline{\varphi}_G$, the projection error can be bounded.

Constructing $\mathcal{D}_{\text{exp}}$ and ρ_{exp}. Therefore, we would like to construct a suitable $\mathcal{D}_{\text{exp}}$ and ρ_{exp} to bound $\|\varphi(s,a)\|_{G^{-1}}$ for any $(s,a) \in \mathcal{S} \times \mathcal{A}$ and solve the following optimization problem:

$$\inf_{\substack{\mathcal{D}_{\text{exp}} \in \mathcal{S} \times \mathcal{A} \\ \rho_{\text{exp}} \in \Delta(\mathcal{D}_{\text{exp}})}} \sup_{(s,a) \in \mathcal{S} \times \mathcal{A}} \|\varphi(s,a)\|_{G^{-1}} \tag{4.15}$$

$$\text{s.t.} \quad G = \sum_{(s,a) \in \mathcal{D}_{\text{exp}}} \rho_{\text{exp}}(s,a) \, \varphi(s,a) \, \varphi(s,a)^\top , \tag{4.16}$$

which fits the form of experimental design. Ideally, we would also like $|\mathcal{D}_{\text{exp}}|$ to be relatively small so that the actor parameters can be updated efficiently.

There are standard techniques to solve this problem. The most common approach is the G-optimal design, which involves constructing a *coreset* (i.e., $\mathcal{D}_{\text{exp}}$ and ρ_{exp}) and bounds $\|\varphi(s,a)\|_{G^{-1}}$. In particular, the Kiefer–Wolfowitz theorem [Kiefer and Wolfowitz, 1960] guarantees that there exists a coreset such that $\|\varphi(s,a)\|_{G^{-1}} \leq \mathcal{O}(d_{\boldsymbol{a}})$ and $|\mathcal{D}_{\text{exp}}| \leq \widetilde{\mathcal{O}}(d_{\boldsymbol{a}})$. Constructing such a coreset can be achieved using various methods, such as the Frank-Wolfe algorithm [Frank et al., 1956, Szepesvári, 2022]. We remark that this method only uses the given policy features φ, and does not involve the linear MDP features ϕ. Furthermore, the required coreset can be constructed offline, even before the learning procedure or without any knowledge of the environment (see Section C.3.1 for details). Giving access to such a coreset guarantees that $\bar{\epsilon} \leq \mathcal{O}(d_{\boldsymbol{a}} \sqrt{\epsilon_{\text{bias}}} + \sqrt{\epsilon_{\text{opt}}})$, and optimizing the actor loss upon that only requires $\mathcal{O}(d_{\boldsymbol{a}})$ computation.

Rather than forming a coreset, alternative approaches assume $\varphi = \phi$, and use some limited interaction with the environment to construct $\mathcal{D}_{\text{exp}}$. In particular, under some standard assumptions (e.g., Wagenmaker and Jamieson, 2022, Assumption 1), we can apply methods such as `CoverTraj` [Wagenmaker et al., 2022] and `OptCov` [Wagenmaker et al., 2022] that construct $\mathcal{D}_{\text{exp}}$ and ρ_{exp} and can subsequently ensure that $\|\varphi(s,a)\|_{G^{-1}}$ is bounded. The sample complexity for such procedures is $\widetilde{\mathcal{O}}(d_{\boldsymbol{c}}^4 H^3 \epsilon_{\mathcal{M}}^{-1})$ where $\epsilon_{\mathcal{M}}$ is a problem-dependent constant. We defer all the details to Section C.3.

Remark 1. *Having access to $\mathcal{D}_{\text{exp}}$ and ρ_{exp} does not obviate the necessity of exploration. Specifically, we do not assume random access, i.e., the agent cannot visit all the state-actions in $\mathcal{D}_{\text{exp}}$. Hence, we cannot effectively calculate $Q^\pi(s,a)$ for any $(s,a) \in \mathcal{D}_{\text{exp}}$.*

4.4.3 Putting Everything Together: Projected NPG with Log-Linear Policies

In Algorithm 3, we instantiate the complete actor algorithm, which uses the projected `NPG` update for log-linear policies. Unlike the standard `NPG` update, Algorithm 3 alleviates the necessity of storing past Q-functions, improving the memory complexity to $\mathcal{O}(d_{\boldsymbol{a}})$, while enjoying similar theoretical guarantees. Furthermore, the actor parameters are updated by using gradient descent on a properly defined surrogate loss, rendering it closer to the practical implementation of common algorithms (e.g., PPO [Schulman et al., 2017b]).

We remark that although we focused on the log-linear policies, our theoretical guarantees readily extend to general function approximation when Assumptions 4 and 5 are satisfied, and one has access to an exploratory policy [Hao et al., 2021, Definition 1]. In the next section, we instantiate the critic in Algorithm 2.

Algorithm 3: Actor: Projected NPG

Input: critic parameters w^t, policy optimization learning rate η, number of actor updates K_t, actor learning rate $\alpha_{\boldsymbol{a}}^t$, subset and distribution of the state-action space $\mathcal{D}_{\text{exp}}$ and ρ_{exp}
for $h = 1, 2, \ldots, H$ **do**
  $\widehat{Q}_h^t(\cdot,\cdot) = \min\{\langle \phi(\cdot,\cdot), w_h^t \rangle, H - h + 1\}^+$ [2]
  Define the actor loss $\widetilde{\ell}_h^t(\theta)$ using Equation (4.13)
  for $k = 1, \ldots, K_t$ **do**
    $\theta_h^{t,k} \leftarrow \theta_h^{t,k-1} - \alpha_{\boldsymbol{a}}^t \nabla_\theta \widetilde{\ell}_h^t(\theta_h^{t,k-1})$
  end for
end for
Return: actor parameters for the policy θ^t

4.5 Instantiating the Critic: Langevin Monte Carlo

In this section, we use Langevin Monte Carlo (LMC) to instantiate the critic. We describe the resulting algorithm in Section 4.5.1, and analyze it in Section 4.5.2

The LMC approaches allow for sampling from a posterior distribution and have recently been used in sequential decision-making problems. For example, Mazumdar et al. [2020] achieves optimal instance-dependent regret bounds for multi-armed bandits using Langevin dynamics for approximate Thompson sampling. On the other hand, Xu et al. [2022] uses LMC for contextual bandits, achieving comparable theoretical results to Thompson sampling. More recently, Ishfaq et al. [2024a] leverages LMC for linear MDPs by using it to sample the Q-function from its posterior distribution, achieving the optimal $\widetilde{\mathcal{O}}(\sqrt{T})$ regret.

Nevertheless, all existing LMC-based approaches for MDPs, including those for general function approximation [Ishfaq et al., 2024b, Jorge et al., 2024] use value-based algorithms. To the best of our knowledge, such approaches have never been theoretically analyzed in the context of policy optimization. Next, we incorporate the LMC algorithm into our actor-critic framework and provide the first provable result.

4.5.1 LMC for Linear MDPs

At episode t, the critic uses the collected dataset $\mathcal{D}^t$ to obtain an optimistic estimate of the Q function. In order to instantiate the critic loss, we consider the dataset $\mathcal{D}^t$ as split into H disjoint subsets $\{\mathcal{D}_h^t\}_{h\in[H]}$, where $\mathcal{D}_h^t$ consists of (s_h, a_h, s_{h+1}) tuples indexed as $\left\{(s_h^i, a_h^i, s_{h+1}^i)\right\}_{i=1}^{|\mathcal{D}^t|}$ [3]. The critic loss at episode t and step h uses the estimated value function

[2] $\min\{x, a\}^+ := \max\{\min\{x, a\}, 0\}$ is the clipping function that bounds the given value x to the range $[0, a]$.

[3] $|\mathcal{D}^t|$ represents the number of trajectories in $\mathcal{D}^t$ or the number of (s_h, a_h, s_{h+1}) tuples in $\mathcal{D}_h^t$

at step $h+1$, and forms the following ridge regression problem:

$$\mathcal{L}_h^t(w) = \tfrac{1}{2} \sum_{i=1}^{|\mathcal{D}^t|} \left[\chi_i^{t,h} - \langle \phi(s_h^i, a_h^i), w \rangle\right]^2 + \tfrac{\lambda}{2} \|w\|^2 , \tag{4.17}$$

$$\text{where } \chi_i^{t,h} = r_h(s_h^i, a_h^i) + \widehat{V}_{h+1}^t(s_{h+1}^i) . \tag{4.18}$$

For each step h, the LMC algorithm iteratively adds Gaussian noise to the gradient descent updates on $\mathcal{L}_h^t(w)$, and aims to produce approximate samples of the critic parameters from its underlying posterior distribution (Line 6-8). In particular, for an arbitrary loss ℓ, the LMC update can be written as:

$$w^{t+1} = w^t - \alpha^t \nabla_w \ell(w^t) + \sqrt{\alpha^t/\zeta}\, \nu^t , \tag{4.19}$$

where α_t is the learning rate, ζ is the inverse temperature parameter, and ν_t is sampled from an isotropic Gaussian distribution. After J_t steps of the LMC update on the critic loss (Lines 6-8 in Algorithm 4), the resulting critic parameters are used to produce an optimistic sample of the Q-function (Line 9). From a theoretical perspective, we note that it is important to clip Q_h^t appropriately. In order to improve the optimism guarantees of the LMC algorithm, we follow the idea in Ishfaq et al. [2021], and repeat the LMC update M times, taking the maximum over these samples (Line 9). Iterating this procedure backwards from $h = H$ to 1, we can obtain the desired critic parameters.

Note that compared to UCB-based approaches, LMC does not require computing confidence sets at every episode. Instead, it simply perturbs gradient descent by injecting Gaussian noise, allowing for a natural extension beyond the linear function approximation setting and rendering it easier to implement in practice.

4.5.2 Optimism Guarantee and Error Bound

In order to theoretically analyze Algorithm 4, we first define the following model prediction error.

Definition 2. *Given an estimated Q-function $\widehat{Q}^t$ and the corresponding estimated value function $\widehat{V}^t$, for all (t, h, s, a), the model prediction error is defined as: $\iota_h^t(s,a) \coloneqq r_h(s,a) + \mathbb{P}_h \widehat{V}_{h+1}^t(s,a) - \widehat{Q}_h^t(s,a)$.*

The theoretical analyses in existing work [Jin et al., 2020, Zhong and Zhang, 2023, Liu et al., 2023] that use UCB bonuses typically proceed by proving an upper bound of 0 on ι_h^t (*optimism*) and a lower bound of $\widetilde{\mathcal{O}}(\sqrt{T})$. The following lemma shows that LMC can offer similar guarantees.

Lemma 2. *Let $\Lambda_h^t \coloneqq \sum_{(s,a,s') \in \mathcal{D}_h^t} \phi(s,a)\, \phi(s,a)^\top + \lambda I$. With appropriate choices of λ, ζ, J_t, α_c^t, M and for any $\delta \in (0,1)$, Algorithm 2 with the LMC critic in Algorithm 4 ensures that in*

Algorithm 4: Critic: LMC

Input: collected data $\mathcal{D}^t$, policy π^{t-1}, number of critic updates J_t, critic learning rate $\alpha_{\boldsymbol{c}}^{h,t}$, inverse temperature ζ, number of critic samples M
$\widehat{V}_{H+1}^t(\cdot) \leftarrow 0$
for $h = H, H-1, \ldots, 1$ **do**
  Define the critic loss $\mathcal{L}_h^t(w)$ using Equation (4.18)
  $w_h^{t,m,0} \leftarrow w_h^{t-1,m,J_{t-1}} \quad \forall m \in [M]$
  for $j = 1, \ldots, J_t$ **do**
    $\nu_h^{t,m,j} \leftarrow \mathbf{N}(0, I) \quad \forall m \in [M]$
    $w_h^{t,m,j} \leftarrow w_h^{t,m,j-1} - \alpha_{\boldsymbol{c}}^{h,t} \nabla_w \mathcal{L}_h^t(w_h^{t,m,j-1}) + \sqrt{\alpha_{\boldsymbol{c}}^{h,t}/\zeta}\, \nu_h^{t,m,j} \quad \forall m \in [M]$
  end for
  $\widehat{Q}_h^t(\cdot,\cdot) = \min\left\{ \max_{m \in [M]} \left\langle \phi(\cdot,\cdot), w_h^{t,m,J_t} \right\rangle, H \right\}^+$
  $\widehat{V}_h^t(\cdot) = \mathbb{E}_{a \sim \pi^{t-1}(\cdot|s)} \widehat{Q}_h^t(\cdot, a)$
end for
Return: critic parameters for the estimated Q-function $\{w_h^{t,m,J_t}\}_{(m,h) \in [M] \times [H]}$

both the on-policy and off-policy settings, for all t, h, s, a and some constant $\Gamma_{\mathtt{LMC}} = \widetilde{\mathcal{O}}(H\, d_{\boldsymbol{c}})$, with probability at least $1 - \delta$,

$$-\Gamma_{\mathtt{LMC}} \times \|\phi(s,a)\|_{(\Lambda_h^t)^{-1}} \leq \iota_h^t(s,a) \leq 0.$$

The exact definition of $\Gamma_{\mathtt{LMC}}$ varies between the on-policy and off-policy settings, although they are both bounded by $\widetilde{O}(H\, d_{\boldsymbol{c}})$ (see Section C.4 for the full version of this lemma). In order to prove this result in the on-policy setting, we use the fact that all the data points in $\mathcal{D}_h^t$ are collected via independent trajectories from the same policy π^t, and are therefore independent and identically distributed. Hence, we can use the self-normalized bounds in Abbasi-Yadkori et al. [2011] to analyze the dependence in h, and prove the corresponding result. In the off-policy setting, since the data points in $\mathcal{D}_h^t$ are collected by different data-dependent policies, these samples are correlated in a complicated manner. Hence, we use the value-aware uniform concentration result from Jin et al. [2020]. We remark that this result requires control over the covering number of the value function class, which is deferred to Section 4.6.

Therefore, we conclude that, compared to UCB bonuses, LMC offers significant practical advantages while still providing similar theoretical guarantees.

4.6 Sample Complexity Analysis

In this section, we analyze the sample complexity of Algorithm 2 with the projected NPG actor from Algorithm 3 and the LMC critic from Algorithm 4. Section 4.6.1 focuses on the on-policy setting, while Section 4.6.2 addresses the off-policy setting.

4.6.1 On-Policy Setting

We now present the following theorem that shows that our proposed algorithm achieves a sample complexity of $\widetilde{\mathcal{O}}(\epsilon^{-4})$ in the on-policy setting.

Theorem 7. *Under Assumptions 4 and 5, consider Algorithm 2 in the on-policy setting with the projected* NPG *actor (Algorithm 3) and the* LMC *critic (Algorithm 4). Suppose $\bar{\epsilon}$ is the projection error in the actor. For an appropriate choice of the actor and critic parameters, including $N = H^4/\epsilon^2$, and any $\delta \in (0,1)$, it holds that with probability at least $1-\delta$,*

$$\mathtt{OG}(T) \leq \widetilde{\mathcal{O}}\left(\frac{H^2\sqrt{d_c^3\,\log|\mathcal{A}|}}{\sqrt{T}} + H^2\sqrt{\bar{\epsilon}}\right). \tag{4.20}$$

Hence, for any $\epsilon > 0$, Algorithm 2 with $T = d_c^3\,H^4\,\log|\mathcal{A}|\,\epsilon^{-2}$ yields a $(\epsilon + H^2\sqrt{\bar{\epsilon}})$-optimal mixture policy, and therefore requires $T \times N = \widetilde{\mathcal{O}}(\epsilon^{-4})$ samples.

Proof Sketch. We decompose the difference between $V_1^{\overline{\pi}^T}(s_1)$ and $V_1^{\star}(s_1)$ into two terms that only depend on either the actor or the critic.

$$\mathbb{E}\left[V_1^{\pi^\star} - V_1^{\overline{\pi}^T}(s_1)\right] = \tag{4.21}$$

$$\underbrace{\frac{1}{T}\sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}_{\pi^\star}\left[\left\langle \pi_h^\star(\cdot \mid s_h) - \pi_h^t(\cdot \mid s_h), \widehat{Q}_h^t(s_h,\cdot)\right\rangle\right]}_{\text{policy optimization (actor) error}} \tag{4.22}$$

$$+\underbrace{\frac{1}{T}\sum_{t=1}^{T}\sum_{h=1}^{H}\left(\mathbb{E}_{\pi^\star}[\iota_h^t(s_h,a_h)] - \mathbb{E}_{\pi^t}[\iota_h^t(s_h,a_h)]\right)}_{\text{policy evaluation (critic) error}}. \tag{4.23}$$

The policy optimization (actor) error can be bounded using Equation (4.10), and the policy evaluation (critic) error is bounded using Lemma 14. In particular, in the on-policy setting, the lower-bound in Lemma 14 can be instantiated as:

$$\sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}_{\pi^t}\left[-\iota_h^t(s,a)\right] \leq \mathcal{O}\left(\sqrt{d_c^3\,H^4\,T\,\log^2(N/\delta)/N}\right). \tag{4.24}$$

Putting everything together with the chosen value of N leads to the stated sample complexity. □

Comparison to Liu et al. [2023]. The on-policy sample-complexity in Theorem 7 matches the bound in Liu et al. [2023, Theorem 1]. However, unlike the proposed algorithm, Liu et al. [2023] uses NPG with implicit policies for the actor. Consequently, sampling from the current policy (to interact with the environment) requires calculating $\sum_{\tau=1}^{t-1} \widehat{Q}_h^\tau(\cdot,\cdot)$ for each encountered state-action pair. Since they use clipped Q functions with `UCB` bonuses for the critic, the above sum of Q functions cannot be stored succinctly. Hence, sampling from the policy requires instantiating each previous Q function for each step h and episode t, resulting in a computational cost of $\mathcal{O}(d_{\boldsymbol{c}}^2\, H\, \epsilon^{-2})$.

4.6.2 Off-Policy Setting

Next, we show that, in the off-policy setting, Algorithm 2 can achieve $\widetilde{\mathcal{O}}(\epsilon^{-2})$ sample complexity.

Theorem 8. *Under Assumptions 4 and 5, consider Algorithm 2 in the off-policy setting with the projected* `NPG` *actor (Algorithm 3) and the* `LMC` *critic (Algorithm 4). Suppose $\overline{\epsilon}$ is the projection error in the actor. For an appropriate choice of the actor and critic parameters and any $\delta \in (0,1)$, it holds that with probability at least $1-\delta$,*

$$\texttt{OG}(T) \leq \widetilde{\mathcal{O}}\left(\frac{H^2\sqrt{d_{\boldsymbol{c}}^3\, \max\{d_{\boldsymbol{a}}, d_{\boldsymbol{c}}\}\, \log|\mathcal{A}|}}{\sqrt{T}} + H^2\,\sqrt{\overline{\epsilon}}\right). \tag{4.25}$$

Hence, for any $\epsilon > 0$, Algorithm 2 with $T = d_{\boldsymbol{c}}^3\, \max\{d_{\boldsymbol{a}}, d_{\boldsymbol{c}}\}\, H^4\, \log|\mathcal{A}|\, \epsilon^{-2}$ yields a $(\epsilon + H^2\sqrt{\overline{\epsilon}})$-optimal mixture policy, and therefore requires $T \times 1 = \widetilde{\mathcal{O}}(\epsilon^{-2})$ samples.

Proof Sketch. The proof uses a similar regret decomposition to Theorem 7. Compared to the on-policy setting, the most significant difference is the bound on the policy evaluation (critic) errors. First, using Lemma 2, we have that

$$\sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}_{\pi^t}\left[-\iota_h^t(s,a)\right] \leq \Gamma \sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}_{\pi^t}\left\|\phi(s,a)\right\|_{(\Lambda_h^t)^{-1}}. \tag{4.26}$$

The term, $\sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}_{\pi^t}\left\|\phi(s,a)\right\|_{(\Lambda_h^t)^{-1}}$, can be bounded using the standard elliptical potential lemma. However, since we are in the off-policy setting, bounding $\Gamma > 0$ requires the uniform concentration argument from Jin et al. [2020], which yields that

$$\Gamma \leq \mathcal{O}\left(H\sqrt{d_{\boldsymbol{c}}\log T + \log\left(\frac{\mathcal{N}_\Delta(\mathcal{V})}{\delta}\right)} + T\,\Delta\right). \tag{4.27}$$

This involves obtaining a bound on $\log\left(\mathcal{N}_\Delta(\mathcal{V})\right)$, the logarithm of the covering number of the value function class. The covering number is a measure of the complexity of the space

of value functions. In particular, we show that for an actor with log-linear policies, we can bound the logarithm of the covering number using the following lemma.

Lemma 3. *Let Π_{lin} be the policy class induced by Equation (4.2) such that $\sup_{\theta,h,s,a} \|z_h(s,a \mid \theta)\| \leq \overline{Z}$. Let $\mathcal{Q} = \left\{\min\left\{\langle \phi(\cdot,\cdot), w\rangle), H\right\}^+ \mid \|w\| \leq \overline{W}\right\}$ be the Q-function class and $\mathcal{V} = \{\langle Q(\cdot,\cdot), \pi(\cdot \mid \cdot,\theta)\rangle_{\mathcal{A}} \mid Q \in \mathcal{Q},\ \pi \in \Pi_{lin}\}$ be the corresponding value function class. Then, it holds that $\log \mathcal{N}_\Delta(\mathcal{V}) \leq \mathsf{V}$ where $\mathsf{V} := d_{\boldsymbol{c}} \log\left(1 + \frac{4\overline{W}+4H\sqrt{2\overline{Z}}}{\Delta}\right) + d_{\boldsymbol{a}} \log\left(1 + \frac{4H\sqrt{2\overline{Z}}}{\Delta}\right)$.*

In particular, we can show that $\overline{W} \leq \mathcal{O}(\sqrt{T})$ (Lemma 19), and $\overline{Z} \leq \mathcal{O}(\bar{\epsilon}\, T)$ (Lemma 34). Putting everything together and setting $\Delta = 1/T$ yields that

$$\sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}_{\pi^t}\left[-\iota_h^t(s,a)\right] \leq \widetilde{\mathcal{O}}\left(\sqrt{d_{\boldsymbol{c}}^3 \max\{d_{\boldsymbol{a}}, d_{\boldsymbol{c}}\}\ H^4\, T}\right). \tag{4.28}$$

Following a proof similar to Theorem 7 leads to the desired sample complexity. □

Remark 2. *In order to effectively bound the logarithm of the covering number, previous work [Sherman et al., 2023, Cassel and Rosenberg, 2024, Tan et al., 2025] has incorporated various algorithmic tweaks, including reward-free warm-ups, feature contractions, and rare-switching. On the contrary, since our algorithm projects the implicit policy onto the log-linear policy class at each iteration, the logarithm of the covering number can be bounded in a more direct manner.*

Comparison to Sherman et al. [2023], Cassel and Rosenberg [2024]. The off-policy sample-complexity in Theorem 8 matches the bound in Sherman et al. [2023], Cassel and Rosenberg [2024]. Similar to Liu et al. [2023], both works use NPG with implicit policies for the actor and `UCB` bonuses for the critic. Consequently, the resulting methods suffer from a high cost of policy inference. Finally, we note that similar to the proposed algorithm (see Section 4.4.2), Sherman et al. [2023] also uses a reward-free warm-up phase [Wagenmaker et al., 2022]. However, while we require the warm-up phase to identify a coreset to efficiently minimize the actor loss (a computational reason), Sherman et al. [2023] requires the warm-up procedure to restrict the subsequent regret minimization procedure to high occupancy regions of the state-action space and effectively control the capacity of the policy class (a statistical reason).

4.7 Experiments

In this section, we first evaluate our proposed algorithm in linear MDPs, consistent with our theoretical analyses. To further demonstrate its versatility, we extend the proposed algorithm to large-scale deep RL applications, evaluating its performance across several Atari games.

4.7.1 Experiments in Linear MDPs

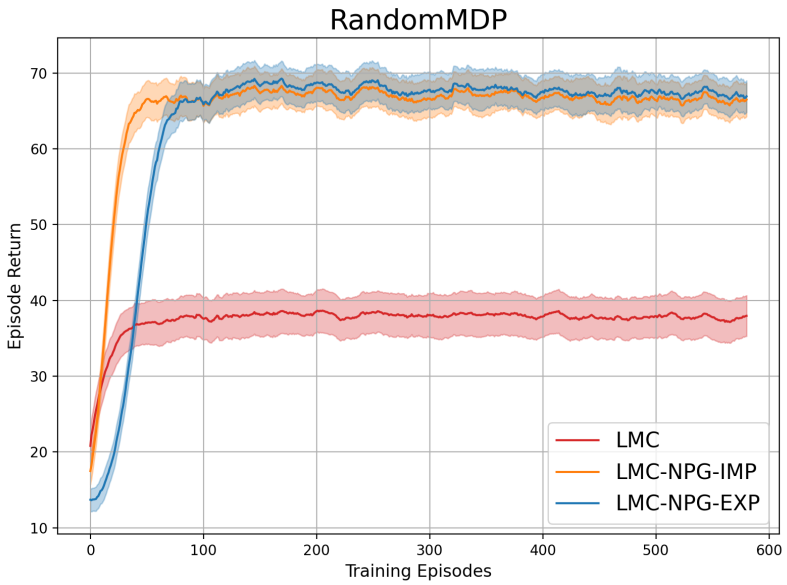


Figure 4.1: Comparison of `LMC-NPG-EXP` (our proposed algorithm), `LMC-NPG-IMP` (memory-intensive variant), and `LMC` (value-based baseline) in the Random MDP.

To demonstrate the practical value of our proposed algorithm, we evaluate it over various benchmarks. We first test our proposed algorithm in the randomly generated linear MDPs (Random MDP). We compare our proposed algorithm, `LMC-NPG-EXP`, with the memory-intensive variant with implicit policy parameterization, `LMC-NPG-IMP`, and the value-based baseline, `LMC` [Ishfaq et al., 2024a]. As shown in Figure 4.1, our proposed algorithm can achieve comparable performance with the memory-intensive variant and better performance than `LMC`. In Section C.7.1, we also do the same experiment in the linear MDP version of the Deep Sea environment [Osband et al., 2019]. In Section C.7.2, we further conduct some ablation studies in these two environments of linear MDPs.

4.7.2 Experiments Beyond Linear MDPs: Atari

We extend our proposed algorithm to large-scale deep RL applications and conduct experiments in Atari [Mnih et al., 2013] using Stable Baselines3 [Raffin et al., 2021]. We compare our proposed algorithm to PPO [Schulman et al., 2017b] in the on-policy setting and DQN [Mnih et al., 2015] in the off-policy setting. The adaptations to deep RL, environment setup, and hyperparameters are described in Appendix C.7.3; the on-policy and off-policy hyperparameters are listed in Tables C.3 and C.4, respectively.

Experimental Results

Figures 4.2 and 4.3 test whether the theoretically motivated LMC-NPG-EXP design remains useful beyond linear MDPs. The method is comparable to or better than PPO in the on-policy setting and matches or exceeds DQN on most games in the off-policy setting. These results support the practical value of the proposed optimistic actor-critic framework for deep RL.

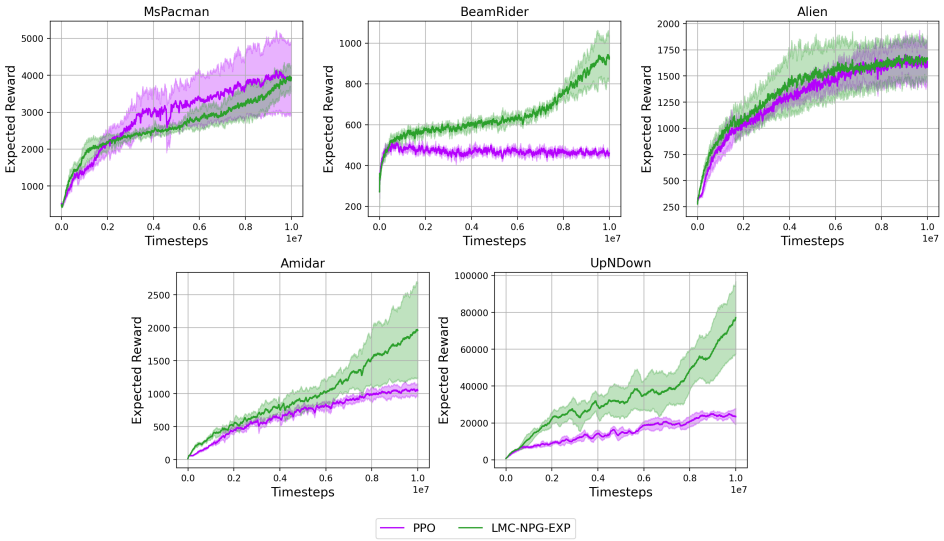


Figure 4.2: In the on-policy setting, `LMC-NPG-EXP` achieves comparable or better performance compared to `PPO`.

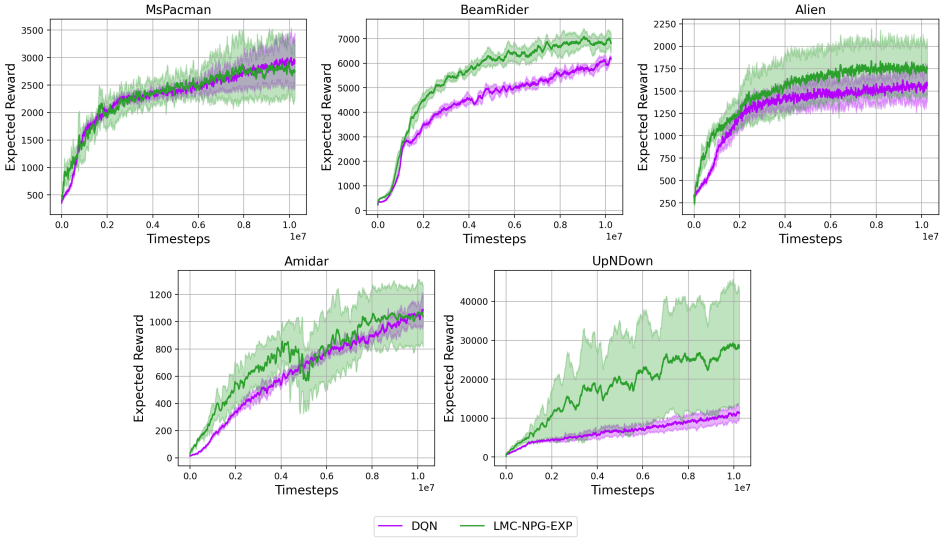


Figure 4.3: In the off-policy setting, `LMC-NPG-EXP` achieves comparable or better performance compared to `DQN`.

4.8 Discussion

We proposed an optimistic actor–critic algorithm with explicitly parameterized policies and a systematic exploration mechanism. In particular, for the actor, we demonstrated that using projected NPG with parametric policies is not only practical, but also equipped with theoretical guarantees. For the critic, we demonstrated that LMC is a principled and easy-to-implement exploration scheme for policy optimization methods. We derived theoretical guarantees in both the on-policy and off-policy settings, showcasing that the proposed actor-critic algorithm can simultaneously achieve sample efficiency and practicality.

For future work, eliminating the dependence on $\mathcal{D}_{\text{exp}}$ and ρ_{exp} will result in a more practical algorithm. We also aim to investigate the actor-critic algorithms in more realistic setups (e.g., infinite-horizon discounted MDPs) with more general function approximation schemes beyond linear models for both the environment and the policy. It would also be fruitful to further evaluate their empirical performance in more challenging large-scale deep RL applications.

Chapter 5

Conclusion

5.1 Summary of Contributions

This thesis studied the design of policy-gradient and actor-critic algorithms that are both theoretically principled and practically effective. The central theme was the interaction between policy optimization, value estimation, data reuse, and exploration. Rather than treating these issues as separate implementation details, the thesis examined how they shape the design and analysis of reinforcement learning algorithms.

The first part of the thesis focused on on-policy policy optimization. Chapter 2 introduced Softmax Policy Mirror Ascent (SPMA), a logit-space mirror-ascent method for policy optimization. It showed that SPMA achieves strong convergence guarantees in tabular and function-approximation settings, while performing competitively on standard deep RL benchmarks such as Atari and MuJoCo without algorithmic modifications.

The second part studied off-policy actor-critic methods for discrete actions. Chapter 3 identified critic entropy regularization as a key bottleneck in discrete SAC and used this observation to develop a modular off-policy actor-critic framework. This framework separates actor entropy from critic entropy and gives rise to new actor objectives, including an off-policy extension of SPMA. The chapter also provided tabular convergence guarantees for the resulting class of objectives and demonstrated their empirical robustness on Atari.

The third part addressed exploration more directly. Chapter 4 studied optimistic actor-critic methods for finite-horizon linear MDPs with explicit parametric policies. The proposed approach combines logit matching for the actor with Langevin-based optimistic value estimation for the critic. This resulted in sample-complexity guarantees in both on-policy and off-policy settings. Experiments in linear MDPs verified the theoretical predictions, while large-scale Atari experiments investigated preliminary practical adaptations of the method.

Together, these results contribute to a broader goal: narrowing the gap between reinforcement learning algorithms that are theoretically analyzable and algorithms that are practically useful. Each part isolates a different aspect of this goal: principled policy optimization, stable

off-policy actor-critic design, and exploration through optimism. The next section discusses how these directions can be extended in future work.

5.2 Future Work

Extending SPMA to preference-based alignment. The SPMA update in Chapter 2 can avoid explicit normalization over the action space under appropriate conditions. This property may be useful in reinforcement learning from human feedback (RLHF), where an action can correspond to an entire response generated by a large language model (LLM). In this setting, normalization over actions requires summing over all possible responses from the LLM, which is intractable. This difficulty is closely related to the normalization term that appears in NPG-based updates. Existing RLHF methods therefore rely on objective reformulations or cancellations to avoid computing such terms [Rafailov et al., 2023, Gao et al., 2024, Azar et al., 2024, Wu et al., 2025, Zhang et al., 2025].

A natural direction is to study whether the SPMA update can be extended to offline preference optimization, leading to an SPMA-style analogue of Direct Preference Optimization (DPO) [Rafailov et al., 2023]. A further direction is to study the corresponding online RLHF setting, where the updated policy is used to collect additional feedback and optimization continues over multiple rounds.

Exploration theory and actor-update efficiency. The analysis in Chapter 3 studies actor-critic updates in the tabular setting, but does not address exploration directly: the critic is obtained through a policy-evaluation procedure over all state-action pairs, so the question of how the data provide sufficient coverage is assumed away. Chapter 4 addresses this missing issue, but only in the linear MDP setting. When adapted to deep RL, the proposed optimistic algorithm does not show a statistically significant improvement over the methods from Chapter 3. This adaptation also requires multiple actor-gradient steps, whereas the methods from Chapter 3 can perform well with a single actor update. In this sense, the methods from Chapter 3 are more efficient in terms of actor updates.

This leaves an important gap between theory and practice. One direction is to extend the theory of Chapter 4 beyond linear MDPs while developing a more actor-efficient version of the algorithm. Another direction is to revisit the analysis of Chapter 3 without sidestepping exploration. Such a study could clarify when actor-critic methods with a single actor-gradient update are sufficient, when optimism provides a more effective exploration mechanism than entropy regularization, and how these mechanisms should be incorporated into practical deep RL algorithms.

Understanding the role of entropy regularization and optimism in deep RL. The experiments in Chapter 4 show that explicit exploration is beneficial in the linear MDP

setting, but the deep RL adaptation does not provide strong evidence for the same conclusion. This is consistent with the results in Chapter 3, where removing entropy regularization from both the actor and the critic, while increasing the number of actor-gradient steps, performs well on many Atari games. A related behavior has been studied theoretically in prior work [Hu et al., 2021].

A useful future direction is therefore to study when entropy regularization and optimism improve deep actor-critic methods, and when comparable gains can instead be obtained by increasing the number of actor-optimization steps. Such a study should go beyond average Atari performance and examine individual games in which these choices lead to different learning dynamics or final returns.

Bibliography

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in Neural Information Processing Systems*, 24, 2011.

Yasin Abbasi-Yadkori, Peter Bartlett, Kush Bhatia, Nevena Lazic, Csaba Szepesvari, and Gellért Weisz. Politex: Regret bounds for policy iteration using expert prediction. In *International Conference on Machine Learning*, pages 3692–3702. PMLR, 2019.

Milton Abramowitz and Irene A Stegun. *Handbook of mathematical functions with formulas, graphs, and mathematical tables*, volume 55. US Government Printing Office, 1948.

Alekh Agarwal, Sham M Kakade, Jason D Lee, and Gaurav Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *J. Mach. Learn. Res.*, 22(98):1–76, 2021a.

Alekh Agarwal, Yujia Jin, and Tong Zhang. VOQL: Towards optimal regret in model-free RL with nonlinear function approximation. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 987–1063. PMLR, 2023.

Naman Agarwal, Syomantak Chaudhuri, Prateek Jain, Dheeraj Nagaraj, and Praneeth Netrapalli. Online target q-learning with reverse experience replay: Efficiently finding the optimal policy for linear mdps. *arXiv preprint arXiv:2110.08440*, 2021b.

Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron C Courville, and Marc Bellemare. Deep reinforcement learning at the edge of the statistical precipice. *Advances in neural information processing systems*, 34:29304–29320, 2021c.

Carlo Alfano and Patrick Rebeschini. Linear convergence for natural policy gradient with log-linear policy parametrization. *arXiv preprint arXiv:2209.15382*, 2022a.

Carlo Alfano and Patrick Rebeschini. Linear convergence for natural policy gradient with log-linear policy parametrization. *arXiv preprint arXiv:2209.15382*, 2022b.

Larry Armijo. Minimization of functions having lipschitz continuous first partial derivatives. *Pacific Journal of mathematics*, 16(1):1–3, 1966.

Reza Asad, Reza Babanezhad, and Sharan Vaswani. Revisiting actor-critic methods in discrete action off-policy reinforcement learning. *arXiv preprint arXiv:2509.09838*, 2025a.

Reza Asad, Reza Babanezhad Harikandeh, Issam H Laradji, Nicolas Le Roux, and Sharan Vaswani. Fast convergence of softmax policy mirror ascent. In *International Conference on Artificial Intelligence and Statistics*, pages 3943–3951. PMLR, 2025b.

Mohammad Gheshlaghi Azar, Zhaohan Daniel Guo, Bilal Piot, Remi Munos, Mark Rowland, Michal Valko, and Daniele Calandriello. A general theoretical paradigm to understand learning from human preferences. In *International Conference on Artificial Intelligence and Statistics*, pages 4447–4455. PMLR, 2024.

Amir Beck and Marc Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.

Marc G Bellemare, Yavar Naddaf, Joel Veness, and Michael Bowling. The arcade learning environment: An evaluation platform for general agents. *Journal of Artificial Intelligence Research*, 47:253–279, 2013.

Marc G Bellemare, Will Dabney, and Rémi Munos. A distributional perspective on reinforcement learning. In *International conference on machine learning*, pages 449–458. PMLR, 2017.

Jalaj Bhandari and Daniel Russo. On the linear convergence of policy gradient methods for finite mdps. In *International Conference on Artificial Intelligence and Statistics*, pages 2386–2394. PMLR, 2021.

Jalaj Bhandari and Daniel Russo. Global optimality guarantees for policy gradient methods. *Operations Research*, 72(5):1906–1927, 2024.

Shalabh Bhatnagar, Richard S Sutton, Mohammad Ghavamzadeh, and Mark Lee. Natural actor–critic algorithms. *Automatica*, 45(11):2471–2482, 2009.

Aditya Bhatt, Daniel Palenicek, Boris Belousov, Max Argus, Artemij Amiranashvili, Thomas Brox, and Jan Peters. Crossq: Batch normalization in deep reinforcement learning for greater sample efficiency and simplicity. *arXiv preprint arXiv:1902.05605*, 2019.

G Brockman. Openai gym. *arXiv preprint arXiv:1606.01540*, 2016.

Sébastien Bubeck et al. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

Qi Cai, Zhuoran Yang, Chi Jin, and Zhaoran Wang. Provably efficient exploration in policy optimization. In *International Conference on Machine Learning*, pages 1283–1294. PMLR, 2020.

Asaf Cassel and Aviv Rosenberg. Warm-up free policy optimization: Improved regret in linear Markov decision processes. *Advances in Neural Information Processing Systems*, 37: 3275–3303, 2024.

Semih Cayci, Niao He, and R. Srikant. Finite-time analysis of entropy-regularized neural natural actor-critic algorithm. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=BkEqk7pS1I.

Shicong Cen, Chen Cheng, Yuxin Chen, Yuting Wei, and Yuejie Chi. Fast global convergence of natural policy gradient methods with entropy regularization. *Operations Research*, 70 (4):2563–2578, 2022.

Nicolo Cesa-Bianchi, Yishay Mansour, and Gilles Stoltz. Improved second-order bounds for prediction with expert advice. *Machine Learning*, 66:321–352, 2007.

Nicolò Cesa-Bianchi, Claudio Gentile, Gábor Lugosi, and Gergely Neu. Boltzmann exploration done right. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/7208-boltzmann-exploration-done-right.pdf.

Alan Chan, Hugo Silva, Sungsu Lim, Tadashi Kozuno, A Rupam Mahmood, and Martha White. Greedification operators for policy optimization: Investigating forward and reverse kl divergences. *Journal of Machine Learning Research*, 23(253):1–79, 2022.

Xinyue Chen, Che Wang, Zijian Zhou, and Keith Ross. Randomized ensembled double q-learning: Learning fast without a model. *arXiv preprint arXiv:2101.05982*, 2021.

Yen-Ju Chen, Nai-Chieh Huang, Ching-pei Lee, and Ping-Chun Hsieh. Accelerated policy gradient: On the convergence rates of the nesterov momentum for reinforcement learning. In *Forty-first International Conference on Machine Learning*, 2023.

Petros Christodoulou. Soft actor-critic for discrete action settings. *arXiv preprint arXiv:1910.07207*, 2019.

Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.

Will Dabney, Mark Rowland, Marc G Bellemare, and Rémi Munos. Distributional reinforcement learning with quantile regression. arxiv e-prints, art. *arXiv preprint arXiv:1710.10044*, 2017.

Will Dabney, Georg Ostrovski, David Silver, and Rémi Munos. Implicit quantile networks for distributional reinforcement learning. In *International conference on machine learning*, pages 1096–1105. PMLR, 2018.

Gabriel Dulac-Arnold, Daniel Mankowitz, and Todd Hester. Challenges of real-world reinforcement learning. *arXiv preprint arXiv:1904.12901*, 2019.

Logan Engstrom, Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Firdaus Janoos, Larry Rudolph, and Aleksander Madry. Implementation matters in deep rl: A case study on ppo and trpo. In *International conference on learning representations*, 2019.

M Fortunato, MG Azar, B Piot, J Menick, I Osband, A Graves, V Mnih, R Munos, D Hassabis, O Pietquin, et al. Noisy networks for exploration. arxiv 2017. *arXiv preprint arXiv:1706.10295*, 2017.

Marguerite Frank, Philip Wolfe, et al. An algorithm for quadratic programming. *Naval research logistics quarterly*, 3(1-2):95–110, 1956.

Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139, 1997.

Zuyue Fu, Zhuoran Yang, and Zhaoran Wang. Single-timescale actor-critic provably finds globally optimal policy. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=pqZV_srUVmK.

Scott Fujimoto, Herke Hoof, and David Meger. Addressing function approximation error in actor-critic methods. In *International conference on machine learning*, pages 1587–1596. PMLR, 2018.

Yaozhong Gan, Renye Yan, Xiaoyang Tan, Zhe Wu, and Junliang Xing. Transductive off-policy proximal policy optimization. *arXiv preprint arXiv:2406.03894*, 2024.

Zhaolin Gao, Jonathan D Chang, Wenhao Zhan, Owen Oertell, Gokul Swamy, Kianté Brantley, Thorsten Joachims, J Andrew Bagnell, Jason D Lee, and Wen Sun. Rebel: Reinforcement learning via regressing relative rewards. *Advances in Neural Information Processing Systems*, 37:52354–52400, 2024.

Mudit Gaur, Amrit Singh Bedi, Di Wang, and Vaneet Aggarwal. On the global convergence of natural actor-critic with two-layer neural network parametrization. *arXiv preprint arXiv:2306.10486*, 2023.

Mudit Gaur, Amrit Bedi, Di Wang, and Vaneet Aggarwal. Closing the gap: Achieving global convergence (last iterate) of actor-critic under Markovian sampling with neural network parametrization. In *International Conference on Machine Learning*, pages 15153–15179. PMLR, 2024.

Evan Greensmith, Peter L Bartlett, and Jonathan Baxter. Variance reduction techniques for gradient estimates in reinforcement learning. *Journal of Machine Learning Research*, 5 (Nov):1471–1530, 2004.

Tuomas Haarnoja, Haoran Tang, Pieter Abbeel, and Sergey Levine. Reinforcement learning with deep energy-based policies. In *International conference on machine learning*, pages 1352–1361. PMLR, 2017.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pages 1861–1870. Pmlr, 2018a.

Tuomas Haarnoja, Aurick Zhou, Kristian Hartikainen, George Tucker, Sehoon Ha, Jie Tan, Vikash Kumar, Henry Zhu, Abhishek Gupta, Pieter Abbeel, et al. Soft actor-critic algorithms and applications. *arXiv preprint arXiv:1812.05905*, 2018b.

Botao Hao, Tor Lattimore, Csaba Szepesvári, and Mengdi Wang. Online sparse reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pages 316–324. PMLR, 2021.

Elad Hazan et al. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325, 2016.

Jiafan He, Heyang Zhao, Dongruo Zhou, and Quanquan Gu. Nearly minimax optimal reinforcement learning for linear markov decision processes. In *International Conference on Machine Learning*, pages 12790–12822. PMLR, 2023.

Matteo Hessel, Joseph Modayil, Hado Van Hasselt, Tom Schaul, Georg Ostrovski, Will Dabney, Dan Horgan, Bilal Piot, Mohammad Azar, and David Silver. Rainbow: Combining improvements in deep reinforcement learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Takuya Hiraoka, Takahisa Imagawa, Taisei Hashimoto, Takashi Onishi, and Yoshimasa Tsuruoka. Dropout q-functions for doubly efficient reinforcement learning. *arXiv preprint arXiv:2110.02034*, 2021.

Mingyi Hong, Hoi-To Wai, Zhaoran Wang, and Zhuoran Yang. A two-timescale stochastic algorithm framework for bilevel optimization: Complexity analysis and application to actor-critic. *SIAM Journal on Optimization*, 33(1):147–180, 2023.

Yuzheng Hu, Ziwei Ji, and Matus Telgarsky. Actor-critic is implicitly biased towards high entropy optimal policies. *arXiv preprint arXiv:2110.11280*, 2021.

Haque Ishfaq, Qiwen Cui, Viet Nguyen, Alex Ayoub, Zhuoran Yang, Zhaoran Wang, Doina Precup, and Lin Yang. Randomized exploration in reinforcement learning with general value function approximation. In *International Conference on Machine Learning*, pages 4607–4616. PMLR, 2021.

Haque Ishfaq, Qingfeng Lan, Pan Xu, A Rupam Mahmood, Doina Precup, Anima Anandkumar, and Kamyar Azizzadenesheli. Provable and practical: Efficient exploration in reinforcement learning via langevin monte carlo. *arXiv preprint arXiv:2305.18246*, 2023.

Haque Ishfaq, Qingfeng Lan, Pan Xu, A Rupam Mahmood, Doina Precup, Anima Anandkumar, and Kamyar Azizzadenesheli. Provable and practical: Efficient exploration in reinforcement learning via langevin monte carlo. In *The Twelfth International Conference on Learning Representations*, 2024a.

Haque Ishfaq, Yixin Tan, Yu Yang, Qingfeng Lan, Jianfeng Lu, A Rupam Mahmood, Doina Precup, and Pan Xu. More efficient randomized exploration for reinforcement learning via approximate sampling. In *Reinforcement Learning Conference*, 2024b.

Haque Ishfaq, Guangyuan Wang, Sami Nur Islam, and Doina Precup. Langevin soft actor-critic: Efficient exploration through uncertainty-driven critic learning. *arXiv preprint arXiv:2501.17827*, 2025.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. In *Conference on learning theory*, pages 2137–2143. PMLR, 2020.

Chi Jin, Qinghua Liu, and Sobhan Miryoosefi. Bellman eluder dimension: New rich classes of rl problems, and sample-efficient algorithms. *Advances in Neural Information Processing Systems*, 34:13406–13418, 2021.

Emmeran Johnson, Ciara Pike-Burke, and Patrick Rebeschini. Optimal convergence rate for exact policy mirror descent in discounted markov decision processes. *arXiv preprint arXiv:2302.11381*, 2023.

Emilio Jorge, Christos Dimitrakakis, and Debabrota Basu. Isoperimetry is all we need: Langevin posterior sampling for rl with sublinear regret. *arXiv preprint arXiv:2412.20824*, 2024.

Sham M Kakade. A natural policy gradient. *Advances in neural information processing systems*, 14, 2001.

Sajad Khodadadian, Prakirt Raj Jhunjhunwala, Sushil Mahavir Varma, and Siva Theja Maguluri. On the linear convergence of natural policy gradient algorithm. In *2021 60th IEEE Conference on Decision and Control (CDC)*, pages 3794–3799. IEEE, 2021.

Sajad Khodadadian, Thinh T Doan, Justin Romberg, and Siva Theja Maguluri. Finite-sample analysis of two-time-scale natural actor–critic algorithm. *IEEE Transactions on Automatic Control*, 68(6):3273–3284, 2022.

Jack Kiefer and Jacob Wolfowitz. The equivalence of two extremum problems. *Canadian Journal of Mathematics*, 12:363–366, 1960.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL `http://arxiv.org`.

Jens Kober, J Andrew Bagnell, and Jan Peters. Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research*, 32(11):1238–1274, 2013.

Vijay Konda and John Tsitsiklis. Actor-critic algorithms. *Advances in Neural Information Processing Systems*, 12, 1999.

Vijay R Konda and John N Tsitsiklis. Actor-critic algorithms. In *Advances in neural information processing systems*, pages 1008–1014, 2000.

Jakub Grudzien Kuba, Christian Schroeder de Witt, and Jakob Foerster. Mirror learning: A unifying framework of policy optimisation. *arXiv preprint arXiv:2201.02373*, 2022.

Arsenii Kuznetsov, Pavel Shvechikov, Alexander Grishin, and Dmitry Vetrov. Controlling overestimation bias with truncated mixture of continuous distributional quantile critics. In *International conference on machine learning*, pages 5556–5566. PMLR, 2020.

Guanghui Lan. Policy mirror descent for reinforcement learning: Linear convergence, new sampling complexity, and generalized problem classes. *Mathematical programming*, 198(1): 1059–1106, 2023.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Jonathan Wilder Lavington, Sharan Vaswani, Reza Babanezhad, Mark Schmidt, and Nicolas Le Roux. Target-based surrogates for stochastic optimization. *arXiv preprint arXiv:2302.02607*, 2023.

Nevena Lazić, Botao Hao, Yasin Abbasi-Yadkori, Dale Schuurmans, and Csaba Szepesvári. Optimization issues in kl-constrained approximate policy iteration. *arXiv preprint arXiv:2102.06234*, 2021.

Yunwen Lei and Yiming Ying. Sharper generalization bounds for learning with gradient-dominated objective functions. In *International Conference on Learning Representations*, 2021.

Gen Li, Yuting Wei, Yuejie Chi, Yuantao Gu, and Yuxin Chen. Softmax policy gradient methods can take exponential time to converge. In *Conference on Learning Theory*, pages 3107–3110. PMLR, 2021.

Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.

Max Qiushi Lin, Reza Asad, Kevin Tan, Haque Ishfaq, Csaba Szepesvari, and Sharan Vaswani. Optimistic actor-critic with parametric policies for linear markov decision processes. *arXiv preprint arXiv:2603.28595*, 2026.

Boyi Liu, Qi Cai, Zhuoran Yang, and Zhaoran Wang. Neural proximal/trust region policy optimization attains globally optimal policy (2019). *arXiv preprint arXiv:1906.10306*, 1906.

Chaoyue Liu, Libin Zhu, and Mikhail Belkin. Loss landscapes and optimization in over-parameterized non-linear systems and neural networks. *Applied and Computational Harmonic Analysis*, 59:85–116, 2022.

Jiacai Liu, Wenye Li, and Ke Wei. Elementary analysis of policy gradient methods. *arXiv preprint arXiv:2404.03372*, 2024.

Qinghua Liu, Gellért Weisz, András György, Chi Jin, and Csaba Szepesvári. Optimistic natural policy gradient: a simple efficient policy optimization framework for online rl. *Advances in Neural Information Processing Systems*, 36:3560–3577, 2023.

Michael Lu, Matin Aghaei, Anant Raj, and Sharan Vaswani. Towards principled, practical policy gradient for bandits and tabular mdps. *arXiv preprint arXiv:2405.13136*, 2024.

Eric Mazumdar, Aldo Pacchiano, Yian Ma, Michael Jordan, and Peter Bartlett. On approximate thompson sampling with langevin algorithms. In *international conference on machine learning*, pages 6797–6807. PMLR, 2020.

Jincheng Mei, Chenjun Xiao, Ruitong Huang, Dale Schuurmans, and Martin Müller. On principled entropy exploration in policy optimization. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 3130–3136, 2019.

Jincheng Mei, Chenjun Xiao, Csaba Szepesvari, and Dale Schuurmans. On the global convergence rates of softmax policy gradient methods. In *International conference on machine learning*, pages 6820–6829. PMLR, 2020.

Jincheng Mei, Bo Dai, Chenjun Xiao, Csaba Szepesvari, and Dale Schuurmans. Understanding the effect of stochasticity in policy optimization. *Advances in Neural Information Processing Systems*, 34:19339–19351, 2021a.

Jincheng Mei, Yue Gao, Bo Dai, Csaba Szepesvari, and Dale Schuurmans. Leveraging non-uniformity in first-order non-convex optimization. In *International Conference on Machine Learning*, pages 7555–7564. PMLR, 2021b.

Wenjia Meng, Qian Zheng, Gang Pan, and Yilong Yin. Off-policy proximal policy optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 9162–9170, 2023.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.

Rémi Munos. Error bounds for approximate value iteration. In *Proceedings of the National Conference on Artificial Intelligence*, volume 20, page 1006. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2005.

Gergely Neu, Anders Jonsson, and Vicenç Gómez. A unified view of entropy-regularized markov decision processes. *arXiv preprint arXiv:1705.07798*, 2017.

Samuel Neumann, Sungsu Lim, Ajin Joseph, Yangchen Pan, Adam White, and Martha White. Greedy actor-critic: A new conditional cross-entropy method for policy improvement. *arXiv preprint arXiv:1810.09103*, 2018.

Samuel Neumann, Jiamin He, Adam White, and Martha White. Investigating the utility of mirror descent in off-policy actor-critic. In *Reinforcement Learning Conference*, 2025.

Konstantinos E Nikolakakis, Farzin Haddadpour, Amin Karbasi, and Dionysios S Kalogerias. Beyond lipschitz: Sharp generalization and excess risk bounds for full-batch gd. *arXiv preprint arXiv:2204.12446*, 2022.

Brendan O'Donoghue. Efficient exploration via epistemic-risk-seeking policy optimization. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 26382–26402. PMLR, 2023. URL `https://proceedings.mlr.press/v202/o-donoghue23a.html`.

Francesco Orabona. A modern introduction to online learning. *arXiv preprint arXiv:1912.13213*, 2019.

Ian Osband, Yotam Doron, Matteo Hessel, John Aslanides, Eren Sezener, Andre Saraiva, Katrina McKinney, Tor Lattimore, Csaba Szepesvari, Satinder Singh, et al. Behaviour suite for reinforcement learning. *arXiv preprint arXiv:1908.03568*, 2019.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Jan Peters, Sethu Vijayakumar, and Stefan Schaal. Natural actor-critic. In *European Conference on Machine Learning*, pages 280–291. Springer, 2005.

Mathieu Poliquin. Stable retro, a maintained fork of openai's gym-retro. `https://github.com/Farama-Foundation/stable-retro`, 2026.

Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

James Queeney, Yannis Paschalidis, and Christos G Cassandras. Generalized proximal policy optimization with sample reuse. *Advances in Neural Information Processing Systems*, 34: 11909–11919, 2021.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

Antonin Raffin. Rl baselines3 zoo. `https://github.com/DLR-RM/rl-baselines3-zoo`, 2020.

Antonin Raffin, Ashley Hill, Adam Gleave, Anssi Kanervisto, Maximilian Ernestus, and Noah Dormann. Stable-baselines3: Reliable reinforcement learning implementations. *Journal of Machine Learning Research*, 22(268):1–8, 2021. URL `http://jmlr.org/papers/v22/20-1364.html`.

Igal Sason. Entropy bounds for discrete random variables via maximal coupling. *IEEE Transactions on Information Theory*, 59(11):7118–7131, 2013.

John Schulman, Sergey Levine, Philipp Moritz, Michael Jordan, and Pieter Abbeel. Trust region policy optimization. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 1889–1897, 2015a. URL `https://proceedings.mlr.press/v37/schulman15.html`.

John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. *arXiv preprint arXiv:1506.02438*, 2015b.

John Schulman, Xi Chen, and Pieter Abbeel. Equivalence between policy gradients and soft q-learning. *arXiv preprint arXiv:1704.06440*, 2017a.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017b.

Lior Shani, Yonathan Efroni, and Shie Mannor. Adaptive trust region policy optimization: Global convergence and faster rates for regularized mdps. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 5668–5675, 2020.

Uri Sherman, Alon Cohen, Tomer Koren, and Yishay Mansour. Rate-optimal policy optimization for linear markov decision processes. *arXiv preprint arXiv:2308.14642*, 2023.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning, Second Edition: An Introduction*. Adaptive Computation and Machine Learning series. The MIT Press, Cambridge, Massachusetts, 2nd edition, 2018. ISBN 978-0-262-03924-6.

Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, 12, 1999.

Csaba Szepesvári. *Algorithms for reinforcement learning*. Springer nature, 2022.

Adrien Ali Taiga, William Fedus, Marlos C Machado, Aaron Courville, and Marc G Bellemare. On bonus-based exploration methods in the arcade learning environment. *arXiv preprint arXiv:2109.11052*, 2021.

Kevin Tan, Wei Fan, and Yuting Wei. Actor-critics can achieve optimal sample efficiency. *arXiv preprint arXiv:2505.03710*, 2025.

Michael J Todd. *Minimum-volume ellipsoids: Theory and algorithms*. SIAM, 2016.

Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 5026–5033. IEEE, 2012. doi: 10.1109/IROS.2012.6386109.

Manan Tomar, Lior Shani, Yonathan Efroni, and Mohammad Ghavamzadeh. Mirror descent policy optimization. *arXiv preprint arXiv:2005.09814*, 2020.

Victor Uc-Cetina, Nicolás Navarro-Guerrero, Anabel Martin-Gonzalez, Cornelius Weber, and Stefan Wermter. Survey on reinforcement learning for language processing. *Artificial Intelligence Review*, 56(2):1543–1575, 2023.

Hado Van Hasselt, Yotam Doron, Florian Strub, Matteo Hessel, Nicolas Sonnerat, and Joseph Modayil. Deep reinforcement learning and the deadly triad. *arXiv preprint arXiv:1812.02648*, 2018.

Sharan Vaswani, Olivier Bachem, Simone Totaro, Robert Müller, Shivam Garg, Matthieu Geist, Marlos C Machado, Pablo Samuel Castro, and Nicolas Le Roux. A general class of surrogate functions for stable and efficient reinforcement learning. *arXiv preprint arXiv:2108.05828*, 2021.

Sharan Vaswani, Amirreza Kazemi, Reza Babanezhad Harikandeh, and Nicolas Le Roux. Decision-aware actor-critic with function approximation and theoretical guarantees. *Advances in Neural Information Processing Systems*, 36, 2024.

Nino Vieillard, Tadashi Kozuno, Bruno Scherrer, Olivier Pietquin, Rémi Munos, and Matthieu Geist. Leverage the average: an analysis of kl regularization in reinforcement learning. *Advances in Neural Information Processing Systems*, 33:12163–12174, 2020a.

Nino Vieillard, Olivier Pietquin, and Matthieu Geist. Munchausen reinforcement learning. *Advances in Neural Information Processing Systems*, 33:4235–4246, 2020b.

Andrew Wagenmaker and Kevin G Jamieson. Instance-dependent near-optimal policy identification in linear mdps via online experiment design. *Advances in Neural Information Processing Systems*, 35:5968–5981, 2022.

Andrew J Wagenmaker, Yifang Chen, Max Simchowitz, Simon Du, and Kevin Jamieson. Reward-free rl is no harder than reward-aware rl in linear markov decision processes. In *International Conference on Machine Learning*, pages 22430–22456. PMLR, 2022.

Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 681–688, 2011.

Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256, 1992.

Yue Wu, Zhiqing Sun, Rina Hughes, Kaixuan Ji, Yiming Yang, and Quanquan Gu. Self-play preference optimization for language model alignment. In *International Conference on Learning Representations*, volume 2025, pages 91558–91582, 2025.

Lin Xiao. On the convergence rates of policy gradient methods. *Journal of Machine Learning Research*, 23(282):1–36, 2022.

Zhihan Xiong, Maryam Fazel, and Lin Xiao. Dual approximation policy optimization. *arXiv preprint arXiv:2410.01249*, 2024.

Pan Xu, Hongkai Zheng, Eric V Mazumdar, Kamyar Azizzadenesheli, and Animashree Anandkumar. Langevin monte carlo for contextual bandits. In *International Conference on Machine Learning*, pages 24830–24850. PMLR, 2022.

Tengyu Xu, Zhe Wang, and Yingbin Liang. Improving sample complexity bounds for (natural) actor-critic algorithms. *Advances in Neural Information Processing Systems*, 33: 4358–4369, 2020.

Tianbing Xu, Qiang Liu, and Jian Peng. Stochastic variance reduction for policy gradient estimation. *arXiv preprint arXiv:1710.06034*, 2017.

Yaosheng Xu, Dailin Hu, Litian Liang, Stephen McAleer, Pieter Abbeel, and Roy Fox. Target entropy annealing for discrete soft actor-critic. *arXiv preprint arXiv:2112.02852*, 2021.

Rui Yuan, Simon S Du, Robert M Gower, Alessandro Lazaric, and Lin Xiao. Linear convergence of natural policy gradient methods with log-linear policies. In *International Conference on Learning Representations*, 2023.

Andrea Zanette, Ching-An Cheng, and Alekh Agarwal. Cautiously optimistic policy optimization and exploration with linear function approximation. In *Conference on Learning Theory*, pages 4473–4525. PMLR, 2021.

Yuheng Zhang, Dian Yu, Baolin Peng, Linfeng Song, Ye Tian, Mingyue Huo, Nan Jiang, Haitao Mi, and Dong Yu. Iterative nash policy optimization: Aligning llms with general preferences via no-regret learning. In *International Conference on Learning Representations*, volume 2025, pages 31833–31849, 2025.

Han Zhong and Tong Zhang. A theoretical analysis of optimistic proximal policy optimization in linear markov decision processes. *Advances in Neural Information Processing Systems*, 36:73666–73690, 2023.

Han Zhong and Tong Zhang. A theoretical analysis of optimistic proximal policy optimization in linear markov decision processes. *Advances in Neural Information Processing Systems*, 36, 2024.

Haibin Zhou, Zichuan Lin, Junyou Li, Qiang Fu, Wei Yang, and Deheng Ye. Revisiting discrete soft actor-critic. *arXiv preprint arXiv:2209.10081*, 2022.

Lingwei Zhu, Haseeb Shah, Han Wang, Yukie Nagai, and Martha White. q-exponential family for policy optimization. *arXiv preprint arXiv:2408.07245*, 2024.

Appendix A

Fast Convergence of Softmax Policy Mirror Ascent

A.1 Multi-armed Bandit Proofs

Theorem 1. *The* SPMA *update in Equation* (2.2) *with (i) a constant step-size* $\eta \leq 1$*, and (ii) uniform initialization i.e.* $\pi_0(a) = \frac{1}{K}$ *for all* a *converges as:*

$$r(a^*) - \langle \pi_T, r \rangle \leq \left(1 - \frac{1}{K}\right) \exp\left(\frac{-\eta\, \Delta_{\min}\, T}{K}\right),$$

where T *is the number of iterations,* a^* *is the optimal arm i.e.* $a^* = \arg\max_a r(a)$ *and* $\Delta_{\min} := \min_{a \neq a^*} \Delta(a^*, a) = r(a^*) - r(a)$ *is the gap.*

Proof. . As in equation (2.2), we can write the update for arm a as following where $\Delta(a, a') = r(a) - r(a')$,

$$\pi_{t+1}(a) = \pi_t(a) \left[1 + \eta \sum_{a' \neq a} \pi_t(a') \Delta(a, a')\right]$$

$$1 - \pi_{t+1}(a^*) = 1 - \pi_t(a^*) - \eta\, \pi_t(a^*) \left[\sum_{a' \neq a^*} \pi_t(a') \Delta(a^*, a')\right] \tag{A.1}$$

We first find a lower-bound for $\sum_{a' \neq a^*} \pi_t(a') \Delta(a^*, a')$:

$$\begin{aligned} \sum_{a' \neq a^*} \pi_t(a') \Delta(a^*, a') &\geq \Delta_{\min} \sum_{a' \neq a^*} \pi_t(a') \\ &= \Delta_{\min}(1 - \pi_t(a^*)) \end{aligned} \tag{A.2}$$

Next, we observe that $\sum_{a' \neq a^*} \pi_t(a') \Delta(a^*, a') \geq 0$. Using this information and starting with a uniform initialization for selecting an arm implies a monotonic improvement on the

probability of selecting the optimal arm:

$$\pi_{t+1}(a^*) > \pi_t(a^*) > ... > \pi_0(a^*) = \frac{1}{K} \tag{A.3}$$

Let $\epsilon_t = 1 - \pi_t(a^*)$.

$$\begin{aligned}
\epsilon_{t+1} &= \epsilon_t - \eta\, \pi_t(a^*) \left[\sum_{a' \neq a^*} \pi_t(a') \Delta(a^*, a') \right] \\
&\leq \epsilon_t - \frac{\eta}{K} \left[\sum_{a' \neq a^*} \pi_t(a') \Delta(a^*, a') \right] && \text{(using (A.3))} \\
&\leq \epsilon_t - \frac{\eta \Delta_{\min}}{K} \epsilon_t && \text{(using (A.2))} \\
&= \epsilon_t \left(1 - \frac{\eta \Delta_{\min}}{K} \right)
\end{aligned}$$

Recursing from $t = 0$ to $t = T - 1$ we have:

$$\begin{aligned}
\epsilon_T &\leq \epsilon_0 \left(1 - \frac{\eta \Delta_{\min}}{K} \right)^T \\
&\leq \epsilon_0 \exp\left(\frac{-\eta \Delta_{\min} T}{K} \right) && \text{(using } 1 - x \leq \exp(-x)\text{)} \\
&= \left(1 - \frac{1}{K} \right) \exp\left(\frac{-\eta \Delta_{\min} T}{K} \right)
\end{aligned}$$

Finally, we define the sub-optimality gap, $\delta_T := r(a^*) - \langle \pi_T, r \rangle$:

$$\begin{aligned}
\delta_T &= \sum_{a'} \pi_T(a') \left[r(a^*) - r(a') \right] \\
&= \sum_{a' \neq a^*} \pi_T(a) \Delta(a^*, a) \\
&\leq \max_{a'} \Delta(a^*, a') \sum_{a' \neq a^*} \pi_T(a) \\
&= \max_{a'} \Delta(a^*, a') (1 - \pi_T(a^*)) \\
&\leq 1 - \pi_T(a^*) && \text{(using the fact } 0 \leq r \leq 1\text{)} \\
&= \epsilon_T \\
&\leq \left(1 - \frac{1}{K} \right) \exp\left(\frac{-\eta \Delta_{\min} T}{K} \right)
\end{aligned}$$

□

A.1.1 Super-linear Rate for Bandits

In order to achieve the desired fast rate of convergence, we modify the update in Equation (2.2) to use a set of $\binom{K}{2}$ constant gap-dependent step-sizes $\{\eta_{a,a'}\}_{a,a' \in [K]}$. The new update can be written as:

$$\pi_{t+1}(a) = \pi_t(a)\,[1 + \sum_{a' \neq a} \pi_t(a')\,\eta_{a,a'}\,\Delta(a, a')] \tag{A.4}$$

The following theorem shows that the above update results in super-linear convergence.

Theorem 9. *Using the* SPMA *update in Equation* (A.4) *with (i)* $\eta_{a,a'} = \frac{1}{|\Delta(a,a')|}$ *and a (ii) uniform initialization similar to Theorem 1 results in valid probability distributions and converges as:*

$$r(a^*) - \langle \pi_T, r \rangle \leq \left[\left(1 - \frac{1}{K}\right)\right]^{2^T}$$

where T *is the number of iterations,* a^* *is the optimal arm and* $\Delta(a, a') := r(a) - r(a')$ *represents the reward gap between arms* a *and* a'*.*

Proof. We define $\Delta(a, a') := r(a) - r(a')$,

$$\begin{aligned} A^{\pi_t} &= r(a) - \langle \pi_t, r \rangle \\ &= \sum_{a'} \pi_t(a')[r(a) - r(a')] \\ &= \sum_{a'} \pi_t(a')\Delta(a, a') \end{aligned}$$

Choosing different step sizes for every pair of arms, depending on their corresponding gap, $\eta_{a,a'} = \frac{1}{|\Delta(a,a')|}$ we get the following update for $\pi_{t+1}(a)$:

$$\begin{aligned} \pi_{t+1}(a) &= \pi_t(a) \left[1 + \sum_{a' \neq a} \eta_{a,a'} \pi_t(a') \Delta(a, a')\right] \\ &= \pi_t(a) \left[1 + \sum_{a' \neq a} \pi_t(a') \operatorname{sign}\left(\Delta(a, a')\right)\right] \end{aligned} \tag{i}$$

Now we check if π_{t+1} is a probability distribution with this choice of η. Note that $\Delta(a, a') = -\Delta(a', a)$.

$$\begin{aligned}
\sum_a \pi_{t+1}(a) &= \sum_a \pi_t(a) + \sum_a \pi_t(a) \sum_{a' \neq a} \pi_t(a') \operatorname{sign}\left(\Delta(a, a')\right) \\
&= 1 + \sum_{(a,a'), a \neq a'} \pi_t(a)\pi_t(a')(\operatorname{sign}\left(\Delta(a, a')\right) + \operatorname{sign}\left(\Delta(a', a)\right)) \\
&= 1 + \sum_{(a,a'), a \neq a'} \pi_t(a)\pi_t(a')(\operatorname{sign}\left(\Delta(a, a')\right) - \operatorname{sign}\left(\Delta(a, a')\right)) \\
&\qquad\qquad (\text{ since } \Delta(a, a') = -\Delta(a', a)) \\
&= 1
\end{aligned}$$

Furthermore, it is clear that $\pi_t(a) \in [0, 1]$. Based on this we just need to show that the probability of the optimal arm a^* converges to 1.

Computing the probability of pulling the optimal arm using update (i):

$$\begin{aligned}
\pi_{t+1}(a^*) &= \pi_t(a^*) \left[1 + \sum_{a' \neq a^*} \pi_t(a') \operatorname{sign}\left(\Delta(a^*, a')\right)\right] \\
&= \pi_t(a^*) \left[1 + \sum_{a' \neq a^*} \pi_t(a')\right] && (\Delta(a^*, a') > 0 \ \ \forall a') \\
&= \pi_t(a^*) \left[2 - \pi_t(a^*)\right] && \text{(ii)}
\end{aligned}$$

We use induction to show $\pi_t(a^*) = 1 - \left[(1 - \frac{1}{K})\right]^{2^t}$ solves the recurrence relation (ii). We consider the uniform distribution over the arms at the initialization i.e. $\pi_0(a) = \frac{1}{K}, \ \ \forall a \in \mathcal{A}$. For the base case, we show the suggested solution satisfies recursion (ii):

$$\begin{aligned}
\pi_1(a^*) &= \frac{1}{K}\left(2 - \frac{1}{K}\right) && \text{(using the recursion in (ii))} \\
&= \left(1 - 1 + \frac{1}{K}\right)\left(1 + 1 - \frac{1}{K}\right) \\
&= 1 - \left[\left(1 - \frac{1}{K}\right)\right]^2
\end{aligned}$$

Assuming the suggested solution is true for t, we show it is also true for $t + 1$:

$$\begin{aligned}
\pi_{t+1}(a^*) &= \left[1 - \left(1 - \frac{1}{K}\right)^{2^t}\right]\left[2 - 1 + \left(1 - \frac{1}{K}\right)^{2^t}\right] \\
&= 1 - \left[\left(1 - \frac{1}{K}\right)^{2^{t+1}}\right]
\end{aligned}$$

Let $\delta_T := r(a^*) - \langle \pi_T, r \rangle$ represent the sub-optimality gap.

$$\begin{aligned}
\delta_T &= \sum_{a'} \pi_T(a') \left[r(a^*) - r(a') \right] \\
&= \sum_{a' \neq a^*} \pi_T(a) \Delta(a^*, a) \\
&\leq \max_{a'} \Delta(a^*, a') \sum_{a' \neq a^*} \pi_T(a) \\
&\leq 1 - \pi_T(a^*) && \text{(using the fact } 0 \leq r \leq 1\text{)} \\
&= \left[\left(1 - \frac{1}{K} \right) \right]^{2^T} && \text{(using the formula for } \pi_T(a^*)\text{)}
\end{aligned}$$

□

A.2 MDP Proofs

A.2.1 Tabular Setting

Lemma 4. *For any policy π_t we have*

$$\sum_a \pi_t(a|s)[A^{\pi_t}(s,a)]^2 \geq \; C_t \max_a A^{\pi_t}(s,a)$$

where $C_t := \min_s \{ \pi_t(\tilde{\mathcal{A}}_t(s)|s) \, \Delta^t(s) \}$, $\tilde{\mathcal{A}}_t(s) := \arg\max_{a \in \mathcal{A}} Q^{\pi_t}(s,a)$, $\pi_t(\tilde{\mathcal{A}}_t(s)|s) = \sum_{a \in \tilde{\mathcal{A}}_t(s)} \pi_t(a_t(s)|s)$ and $\Delta^t(s) := \max_{a \in \mathcal{A}} Q^{\pi_t}(s,a) - \max_{a \notin \tilde{\mathcal{A}}} Q^{\pi_t}(s,a)$.

Proof. Recall $\tilde{\mathcal{A}}_t(s) := \arg\max_{a \in \mathcal{A}} A^{\pi_t}(s,a)$ i.e. $\tilde{\mathcal{A}}_t(s)$ is a set containing actions with maximum advantage for state s. Let's define $\pi_t(\tilde{\mathcal{A}}_t(s)|s) = \sum_{a \in \tilde{\mathcal{A}}_t(s)} \pi_t(\tilde{a}_t(s)|s)$. We can split the sum on the LHS of the above over $\tilde{\mathcal{A}}_t(s)$:

$$\begin{aligned}
\sum_a \pi_t(a|s)[A^{\pi_t}(s,a)]^2 &= \sum_{a \in \tilde{\mathcal{A}}_t(s)} \pi_t(\tilde{a}_t(s)|s)[\max_a A^{\pi_t}(s,a)][\max_a A^{\pi_t}(s,a)] \\
&\quad + \sum_{a \notin \tilde{\mathcal{A}}_t(s)} \pi_t(a|s)[A^{\pi_t}(s,a)]^2 \\
&= \pi_t(\tilde{\mathcal{A}}_t(s)|s)[\max_a A^{\pi_t}(s,a)][\max_a A^{\pi_t}(s,a)] \\
&\quad + \sum_{a \notin \tilde{\mathcal{A}}_t(s)} \pi_t(a|s)[A^{\pi_t}(s,a)]^2
\end{aligned} \tag{A.5}$$

Let $\tilde{\pi}_t$ be the following distribution over the actions.

$$\tilde{\pi}_t(a|s) = \begin{cases} 0 & \text{if } a \in \tilde{\mathcal{A}}_t(s) \\ \frac{\pi_t(a|s)}{1 - \pi_t(\tilde{\mathcal{A}}_t(s)|s)} & \text{otherwise} \end{cases}$$

Re-writing $\sum_a \pi_t(a|s)A^{\pi_t}(s,a) = 0$ using the above distribution we obtain:

$$(1 - \pi_t(\tilde{\mathcal{A}}_t(s)|s))\ \mathbb{E}_{a\sim\tilde{\pi}_t}[A^{\pi_t}(s,a)] + \pi_t(\tilde{\mathcal{A}}_t(s)|s)[\max_a A^{\pi_t}(s,a)] = 0$$

$$(1 - \pi_t(\tilde{\mathcal{A}}_t(s)|s))\ \mathbb{E}_{a\sim\tilde{\pi}_t}[A^{\pi_t}(s,a)] = -\pi_t(\tilde{\mathcal{A}}_t(s)|s)[\max_a A^{\pi_t}(s,a)] \tag{A.6}$$

Expanding the second term in Eq. A.5 using $\tilde{\pi}_t$ we obtain:

$$\begin{aligned}
\sum_{a\notin\tilde{\mathcal{A}}_t(s)} \pi_t(a|s)[A^{\pi_t}(s,a)]^2 =&\ (1 - \pi_t(\tilde{\mathcal{A}}_t(s)|s))\ \mathbb{E}_{a\sim\tilde{\pi}_t}[A^{\pi_t}(s,a)]^2 \\
\geq&\ (1 - \pi_t(\tilde{\mathcal{A}}_t(s)|s))\ (\mathbb{E}_{a\sim\tilde{\pi}_t}[A^{\pi_t}(s,a)])^2 \quad (\text{using } \mathbb{E}[x^2] \geq (\mathbb{E}[x])^2) \\
=&\ (1 - \pi_t(\tilde{\mathcal{A}}_t(s)|s))\ (\mathbb{E}_{a\sim\tilde{\pi}_t}[A^{\pi_t}(s,a)])\,(\mathbb{E}_{a\sim\tilde{\pi}_t}[A^{\pi_t}(s,a)]) \\
=&-\pi_t(\tilde{\mathcal{A}}_t(s)|s)[\max_a A^{\pi_t}(s,a)]\,(\mathbb{E}_{a\sim\tilde{\pi}_t}[A^{\pi_t}(s,a)])\,(\text{using Eq. A.6})
\end{aligned}$$

Plugging in the result above into Eq. A.5 we obtain:

$$\begin{aligned}
\sum_a \pi_t(a|s)[A^{\pi_t}(s,a)]^2 \geq&\ \pi_t(\tilde{\mathcal{A}}_t(s)|s)[\max_a A^{\pi_t}(s,a)][\max_a A^{\pi_t}(s,a)] \\
-&\ \pi_t(\tilde{\mathcal{A}}_t(s)|s)[\max_a A^{\pi_t}(s,a)]\,(\mathbb{E}_{a\sim\tilde{\pi}_t}[A^{\pi_t}(s,a)]) \\
\geq&\ \pi_t(\tilde{\mathcal{A}}_t(s)|s)[\max_a A^{\pi_t}(s,a)]\left[\max_a A^{\pi_t}(s,a) - \mathbb{E}_{a\sim\tilde{\pi}_t}[A^{\pi_t}(s,a)]\right] \\
\geq&\ \pi_t(\tilde{\mathcal{A}}_t(s)|s)[\max_a A^{\pi_t}(s,a)]\left[\underbrace{\max_a A^{\pi_t}(s,a) - \max_{a\notin\tilde{\mathcal{A}}} A^{\pi_t}(s,a)}_{:=\Delta^t(s)}\right] \\
=&\ \pi_t(\tilde{\mathcal{A}}_t(s)|s)[\max_a A^{\pi_t}(s,a)]\Delta^t(s) \\
\geq&\ C_t \max_a A^{\pi_t}(s,a)
\end{aligned}$$

□

Lemma 5. *Using the update $\pi_{t+1}(a|s) = \pi_t(a|s)\,(1 + \eta A^{\pi_t}(s,a))$ with a step-size $\eta < \min\left\{1-\gamma, \frac{1}{C_t(1-\gamma)}\right\}$, at any iteration t and state $s \in S$, we have*

$$V^*(s) - V^{\pi_{t+1}}(s) \leq [1 - \eta\ C_t(1-\gamma)]\,[V^*(s) - V^{\pi_t}(s)]$$

where $C_t := \min_s\{\pi_t(\tilde{\mathcal{A}}_t(s)|s)\,\Delta^t(s)\}$, $\tilde{\mathcal{A}}_t(s) := \arg\max_a Q^{\pi_t}(s,a)$, $\pi_t(\tilde{\mathcal{A}}_t(s)|s) = \sum_{a\in\tilde{\mathcal{A}}_t(s)} \pi_t(a_t(s)|s)$, $\Delta^t(s) := \max_a Q^{\pi_t}(s,a) - \max_{a\notin\tilde{\mathcal{A}}} Q^{\pi_t}(s,a)$, and $V^*(s)$ is the value function corresponding to the optimal policy π^* at $s \in S$.

Proof. First, we use the value difference Lemma to show the SPMA update in Equation (2.3) leads to a monotonic improvement in the value function.

$$V^{\pi_{t+1}}(s) - V^{\pi_t}(s) = \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\pi_{t+1}}}\left[\sum_a \pi_{t+1}(a|s)A^{\pi_t}(s,a)\right] \tag{A.7}$$

Plugging update Equation (2.3) into the term within the brackets, we obtain the following:

$$\begin{aligned}
\sum_a \pi_{t+1}(a|s)A^{\pi_t}(s,a) &= \sum_a \pi_t(a|s)A^{\pi_t}(s,a)[1+\eta A^{\pi_t}(s,a)] \\
&= \sum_a \pi_t(a|s)A^{\pi_t}(s,a) + \eta\sum_a \pi_t(a|s)[A^{\pi_t}(s,a)]^2 \\
&= \eta\sum_a \pi_t(a|s)[A^{\pi_t}(s,a)]^2 \\
&> 0
\end{aligned}$$

Hence, $V^{\pi_{t+1}}(s) \geq V^{\pi_t}(s)$. Using Lemma 4, we have:

$$\eta\sum_a \pi_t(a|s)[A^{\pi_t}(s,a)]^2 \geq \eta\ C_t \max_a A^{\pi_t}(s,a) \tag{A.8}$$

Combining the above with the result from the value difference Lemma we have:

$$\begin{aligned}
\sum_a \pi_{t+1}(a|s)A^{\pi_t}(s,a) &= \eta\sum_a \pi_t(a|s)[A^{\pi_t}(s,a)]^2 \\
&\geq \eta\ C_t \max_a A^{\pi_t}(s,a)
\end{aligned} \tag{A.9}$$

We now show a linear convergence when using the update in Equation (2.3). Let T be the Bellman optimality operator defined as:

$$(Tv)(s) = \max_a\{r(s,a) + \gamma\sum_{s'}\Pr[s'|s,a]v(s')\}$$

Applying the operator at iteration t we obtain:

$$TV^{\pi_t}(s) - V^{\pi_t}(s) = \max_a Q^{\pi_t}(s,a) - V^{\pi_t}(s) = \max_a A^{\pi_t}(s,a) \tag{A.10}$$

Let T^π be an operator w.r.t π defined as:

$$T^\pi(v) = \sum_a \pi(a|s)r(s,a) + \gamma\sum_a \pi(a|s)\sum_{s'}\Pr[s'|s,a]v(s')$$

Applying T^π to $V^{\pi'}(s)$ results in:

$$\begin{aligned}
T^\pi V^{\pi'}(s) &= \sum_a \pi(a|s)r(s,a) + \gamma\sum_a \pi(a|s)\sum_{s'}\Pr[s'|s,a]V^{\pi'}(s) \\
&= \sum_a \pi(a|s)Q^{\pi'}(s,a)
\end{aligned}$$

Using the above we obtain:

$$\begin{aligned}
T^{\pi_{t+1}}V^{\pi_t}(s) - V^{\pi_t}(s) &= \sum_a \pi_{t+1}(a|s)A^{\pi_t}(s,a) \\
&\geq \eta\ C_t \max_a A^{\pi_t}(s,a) && \text{(using Ineq. A.9)} \\
&= \eta\ C_t\left[TV^{\pi_t}(s) - V^{\pi_t}(s)\right] && \text{(using Eq. A.10)}
\end{aligned}$$

Assuming π^* is the optimal policy we have:

$$\begin{aligned}
V^*(s) - V^{\pi_{t+1}}(s) &= V^*(s) - T^{\pi_{t+1}}V^{\pi_{t+1}}(s) && \text{(since } T^\pi V^\pi(s) = V^\pi(s)\text{)} \\
&\leq V^*(s) - T^{\pi_{t+1}}V^{\pi_t}(s) && \text{(since } V^{\pi_{t+1}}(s) \geq V^{\pi_t}(s)\ \ \forall s\text{)} \\
&= V^*(s) - V^{\pi_t}(s) - [T^{\pi_{t+1}}V^{\pi_t}(s) - V^{\pi_t}(s)] && \text{(add and subtract } V^{\pi_t}(s)\text{)} \\
&\leq V^*(s) - V^{\pi_t}(s) - \eta\ C_t\left[TV^{\pi_t}(s) - V^{\pi_t}(s)\right] \\
&= \eta\ C_t[V^*(s) - V^{\pi_t}(s)] + (1 - \eta\ C_t)[V^*(s) - V^{\pi_t}(s)] - \eta\ C_t\left[TV^{\pi_t}(s) - V^{\pi_t}(s)\right] \\
&= \eta\ C_t\left[TV^*(s) - V^{\pi_t}(s) - TV^{\pi_t}(s) + V^{\pi_t}(s)\right] + (1 - \eta\ C_t)\left[V^*(s) - V^{\pi_t}(s)\right] \\
&= \eta\ C_t\left[TV^*(s) - TV^{\pi_t}(s)\right] + (1 - \eta\ C_t)\left[V^*(s) - V^{\pi_t}(s)\right] \\
&\leq \gamma\ \eta\ C_t\left[V^*(s) - V^{\pi_t}(s)\right] + (1 - \eta\ C_t)\left[V^*(s) - V^{\pi_t}(s)\right] \\
& && (T \text{ is a } \gamma \text{ contraction map}) \\
&= \left[1 - \eta\ C_t(1-\gamma)\right]\left[V^*(s) - V^{\pi_t}(s)\right]
\end{aligned}$$

□

Theorem 2. *Using the* SPMA *update in Equation* (2.3) *with a step-size* $\eta < \min\left\{1-\gamma, \frac{1}{C_t(1-\gamma)}\right\}$ *converges as:*

$$\left\|V^{\pi^*} - V^{\pi_T}\right\|_\infty \leq \left(\prod_{t=0}^{T-1} \alpha_t\right)\left\|V^{\pi^*} - V^{\pi_0}\right\|_\infty,$$

where $\alpha_t := (1 - \eta\, C_t\,(1-\gamma))$, $C_t := \min_s\{\pi_t(\tilde{a}_t(s)|s)\,\Delta^t(s)\}$, $\tilde{a}_t(s) := \arg\max_a Q^{\pi_t}(s,a)$ *and* $\Delta^t(s) := \max_a Q^{\pi_t}(s,a) - \max_{a\neq\tilde{a}} Q^{\pi_t}(s,a)$.

Proof. Using Lemma 5 we have

$$V^*(s) - V^{\pi_{t+1}}(s) \leq \left[1 - \eta\ C_t(1-\gamma)\right]\left[V^*(s) - V^{\pi_t}(s)\right]$$

If $\eta < \frac{1}{C_t(1-\gamma)}$, both sides of the inequality above are positive leading to $|V^*(s) - V^{\pi_{t+1}}(s)| \leq (1 - \eta\ C_t(1-\gamma))|V^*(s) - V^{\pi_t}(s)|$. This is true for all $s \in \mathcal{S}$, hence we have:

$$\begin{aligned}
\left\|V^{\pi^*} - V^{\pi_{t+1}}\right\|_\infty &\leq (1 - \eta\ C_t(1-\gamma))\left\|V^{\pi^*} - V^{\pi_t}\right\|_\infty \\
&= \alpha_t\left\|V^{\pi^*} - V^{\pi_t}\right\|_\infty
\end{aligned}$$

Recursing from $t = 0$ to $t = T-1$ we obtain a linear convergence:

$$\left\|V^{\pi^*} - V^{\pi_T}\right\|_\infty \leq \left(\prod_{t=0}^{T-1} \alpha_t\right)\left\|V^{\pi^*} - V^{\pi_0}\right\|_\infty$$

□

A.2.2 Function Approximation With Exact Advantage

Recall the definitions of $\tilde{\ell}_t$ and ℓ_t

$$\tilde{\ell}_t(\theta) = \sum_s d^{\pi_t}(s)\,\mathrm{KL}(\pi_{t+1/2}(\cdot|s)\,||\,\pi_\theta(\cdot|s))$$
$$\ell_t(\theta) = \sum_{s\sim\tau} \mathrm{KL}\left(h(f_{\theta_t}(s,\cdot))(1+\eta A^{\pi_t}(s,\cdot))\,||\,h(f_\theta(s,\cdot))\right)$$

Theorem 3. *Under Assumption 1-3,* SPMA *with* $\eta < \min\left\{1-\gamma, \frac{1}{C_t(1-\gamma)}\right\}$ *converges as,*

$$J(\pi^*) - J(\pi_T)$$
$$\leq \left(\prod_{t=0}^{T-1}\alpha_t\right)(J(\pi^*) - J(\pi_0)) + \beta\sum_{t=0}^{T-1}\prod_{i=t+1}^{T-1}\alpha_i\,,$$

where $\beta = \frac{\sqrt{2}}{(1-\gamma)^2\rho_{\min}}\sqrt{\epsilon_{stat}+\epsilon_{bias}}$ *and* α_t *has the same definition as in Theorem 2.*

Proof. We assumed that $z_\theta(s,a) := f_\theta(s,a)\ \forall(s,a)$ and $z_t(s,a) = f_{\theta_{t+1}}(s,a)$ where $f_\theta : \mathbb{R}^{SA} \to \mathbb{R}$ is a complex, non-linear model. We remind the following updates:

$$z_{t+1/2} = \arg\max_{\bar{z}\in\mathbb{R}^{|S||A|}}\{\langle\nabla_z J(z_t), z\rangle - 1/\eta D_\Phi(z, z_t)\}$$
$$\nabla\Phi(z_{t+1/2}) = \nabla\Phi(z_t) + \eta\nabla_z J(z_t) \qquad \text{(Mirror Ascent update without projection)}$$
$$\pi_{t+1/2} = h(z_{t+1/2}) \qquad (h \text{ is softmax})$$
$$\pi_{t+1/2}(a|s) = \pi_t(a|s)(1+\eta A^{\pi_t}(s,a))$$
$$\theta_{t+1} = \text{(S)GD}(\ell_t(\theta)) \quad \text{(using (Stochastic)Gradient Descent for } m \text{ iteration to minimize } \ell_t)$$
$$\pi_{t+1} = h(z_{t+1})$$

$z_{t+1/2}$ is an unprojected update for the tabular setting and therefore using Lemma 5 we have:

$$V^*(s) - V^{\pi_{t+1/2}}(s) \leq [1-\eta\ C_t(1-\gamma)]\,[V^*(s) - V^{\pi_t}(s)]$$

By adding and removing $V^{\pi_{t+1}}(s)$ to both sides and rearranging we have

$$V^*(s) - V^{\pi_{t+1}}(s) \leq [1-\eta\ C_t(1-\gamma)]\,[V^*(s) - V^{\pi_t}(s)] + V^{\pi_{t+1/2}}(s) - V^{\pi_{t+1}}(s).$$

Taking the expectation w.r.t. ρ, we obtain:

$$J(\pi^*) - J(\pi_{t+1}) \leq [1 - \eta\ C_t(1-\gamma)]\,[J(\pi^*) - J(\pi_t)] + \underbrace{J(\pi_{t+1/2}) - J(\pi_{t+1})}_{:=E_1}$$

The term E_1 can be bounded as follows:

$$\begin{aligned}
E_1 &= \sum_s d^{\pi_{t+1/2}}(s)\langle Q^{\pi_{t+1}}(s,.), \pi_{t+1/2}(.|s) - \pi_{t+1}(.|s)\rangle \\
&\leq \sum_s d^{\pi_{t+1/2}}(s)\, \|Q^{\pi_{t+1}}(s,.)\|_\infty \left\|\pi_{t+1/2}(.|s) - \pi_{t+1}(.|s)\right\|_1 && \text{(Holder inequality)} \\
&\leq \frac{1}{1-\gamma}\sum_s d^{\pi_{t+1/2}}(s) \left\|\pi_{t+1/2}(.|s) - \pi_{t+1}(.|s)\right\|_1 && \text{(since } \|Q^{\pi_{t+1}}(s,.)\|_\infty \leq \tfrac{1}{1-\gamma}\text{)} \\
&= \frac{1}{1-\gamma}\sum_s \frac{d^{\pi_{t+1/2}}(s)}{\rho(s)}\rho(s) \left\|\pi_{t+1/2}(.|s) - \pi_{t+1}(.|s)\right\|_1 \\
&\leq \frac{1}{1-\gamma}\left\|\frac{d^{\pi_{t+1/2}}}{\rho}\right\|_\infty \sum_s \rho(s) \left\|\pi_{t+1/2}(.|s) - \pi_{t+1}(.|s)\right\|_1 \\
&\leq \frac{1}{(1-\gamma)\rho_{min}}\sum_s \rho(s) \left\|\pi_{t+1/2}(.|s) - \pi_{t+1}(.|s)\right\|_1 \\
& && \text{(since } d^{\pi_{t+1/2}}(s) \leq 1 \text{ and using Assumption 3)} \\
&\leq \frac{1}{(1-\gamma)^2\rho_{min}}\sum_s d^{\pi_t}(s) \left\|\pi_{t+1/2}(.|s) - \pi_{t+1}(.|s)\right\|_1 && \text{(sicne } d^{\pi_t} \geq (1-\gamma)\rho\text{)} \\
&\leq \frac{\sqrt{2}}{(1-\gamma)^2\rho_{min}}\sum_s d^{\pi_t}(s)\sqrt{\mathrm{KL}(\pi_{t+1/2}(.|s)\,||\,\pi_{t+1}(.|s))} \\
& && \text{(using strong convexity of KL divergence or Pinsker's inequality)} \\
&\leq \frac{\sqrt{2}}{(1-\gamma)^2\rho_{min}}\sqrt{\underbrace{\sum_s d^{\pi_t}(s)\mathrm{KL}(\pi_{t+1/2}(.|s)\,||\,\pi_{t+1}(.|s))}_{:=E_2}} \\
& && \text{(due to concavity of } \sqrt{} \text{ and Jensen's inequality)}
\end{aligned}$$

where E_2 can be bounded as follows.

$$\begin{aligned}
E_2 &= \tilde{\ell}_t(\theta_{t+1}) \\
&= \tilde{\ell}_t(\theta_{t+1}) - \min_\theta \tilde{\ell}_t(\theta_{t+1}) + \min_\theta \tilde{\ell}_t(\theta_{t+1}) \\
&\leq \epsilon_{stat} + \min_\theta \tilde{\ell}_t(\theta_{t+1}) && \text{(using Assumption 1)} \\
&\leq \epsilon_{stat} + \epsilon_{bias} && \text{(using Assumption 2)}
\end{aligned}$$

Putting everything together we have:

$$E_1 \leq \underbrace{\frac{\sqrt{2}}{(1-\gamma)^2\rho_{min}}\sqrt{\epsilon_{\text{stat}} + \epsilon_{\text{bias}}}}_{:=\beta}$$

Therefore we have

$$J(\pi^*) - J(\pi_{t+1}) \leq \underbrace{[1 - \eta\ C_t(1-\gamma)]}_{\alpha_t} [J(\pi^*) - J(\pi_t)] + \beta$$

Unrolling the above recursion for T iterations,

$$J(\pi^*) - J(\pi_T) \leq \left(\prod_{t=0}^{T-1} \alpha_t\right) (J(\pi^*) - J(\pi_0)) + \beta \sum_{t=0}^{T-1} \prod_{i=t+1}^{T-1} \alpha_t$$

□

A.2.3 Function Approximation With Inexact Advantage

In practice computing the A^{π_t} at every iteration is costly. In this section, we assume that we access an oracle such that at each iteration t, it gives us $\hat{A}^{\pi_t}$ an approximation of A^{π_t}.

Assumption 6. *Valid Approximation. For all iterations t and $(s,a) \in \mathcal{S} \times \mathcal{A}$, $|\hat{A}^{\pi_t}(s,a)| \leq \frac{1}{1-\gamma}$.*

Assumption 7. *Approximation Error. For all iterations t and $s \in \mathcal{S}$, $\left\|A^{\pi_t}(s,.) - \hat{A}^{\pi_t}(s,.)\right\|_\infty \leq \epsilon_{approx}$.*

Using this inexact advantage function, we define the following update and functions

$$\begin{aligned}
&\pi_{t+1/2}(a|s) = \pi_t(a|s)(1 + \eta \hat{A}^{\pi_t}(s,a)) && \text{(replacing } A^{\pi_t} \text{ with } \hat{A}^{\pi_t}\text{)} \\
&\tilde{\ell}_t(\theta) = \sum_s d^{\pi_t}(s)\, \mathrm{KL}(\pi_{t+1/2}(\cdot|s)\,||\,\pi_\theta(\cdot|s)) \\
&\ell_t(\theta) = \sum_{s\sim\tau} \mathrm{KL}\left(h(f_{\theta_t}(s,\cdot))(1 + \eta \hat{A}^{\pi_t}(s,\cdot))\,||\,h(f_\theta(s,\cdot))\right)
\end{aligned}$$

Since we use the inexact advantage, we cannot reuse the result of Lemma 5. So we provide a variant of that lemma with an inexact advantage.

Lemma 6. *Using the update $\pi_{t+1}(a|s) = \pi_t(a|s)\,(1 + \eta \hat{A}^{\pi_t}(s,a))$ with (i) a step-size $\eta < \min\left\{1-\gamma, \frac{1}{C_t(1-\gamma)}\right\}$ and (ii) $\hat{A}^{\pi_t}$ satisfying assumptions 6 and 7, at any iteration t and $s \in S$ we have*

$$V^*(s) - V^\pi(s) \leq [1 - \eta\ C_t(1-\gamma)]\,[V^*(s) - V^{\pi_t}(s)] + \frac{\epsilon_{approx}}{1-\gamma}$$

where $C_t := \min_s\{\pi_t(\tilde{\mathcal{A}}_t(s)|s)\,\Delta^t(s)\}$, $\tilde{\mathcal{A}}_t(s) := \arg\max_a Q^{\pi_t}(s,a)$, $\pi_t(\tilde{\mathcal{A}}_t(s)|s) = \sum_{a\in\tilde{\mathcal{A}}_t(s)} \pi_t(a_t(s)|s)$ and $\Delta^t(s) := \max_a Q^{\pi_t}(s,a) - \max_{a\notin\tilde{\mathcal{A}}} Q^{\pi_t}(s,a)$, and $V^*(s)$ is the value function corresponding to the optimal policy π^* at $s \in \mathcal{S}$.

Proof. First, we use the value difference Lemma for a state $s \in \mathcal{S}$

$$
\begin{aligned}
V^{\pi}(s) - V^{\pi_t}(s) &= \frac{1}{1-\gamma}\mathbb{E}_{s'\sim d^{\pi}}\left[\sum_a \pi(a|s')A^{\pi_t}(s',a)\right] \\
&= \frac{1}{1-\gamma}\mathbb{E}_{s'\sim d^{\pi}}\left[\sum_a \pi_t(a|s')\left(1+\eta\hat{A}^{\pi_t}(s',a)\right)A^{\pi_t}(s',a)\right] \\
&= \frac{1}{1-\gamma}\mathbb{E}_{s'\sim d^{\pi}}\left[\sum_a \eta\pi_t(a|s')\,\hat{A}^{\pi_t}(s',a)A^{\pi_t}(s',a)\right] \\
&= \frac{1}{1-\gamma}\mathbb{E}_{s'\sim d^{\pi}}\left[\sum_a \eta\pi_t(a|s')\left(\hat{A}^{\pi_t}(s',a) - A^{\pi_t}(s',a) + A^{\pi_t}(s',a)\right)A^{\pi_t}(s',a)\right] \\
&= \frac{1}{1-\gamma}\mathbb{E}_{s'\sim d^{\pi}}\underbrace{\left[\sum_a \eta\pi_t(a|s')\left(A^{\pi_t}(s',a)\right)^2\right]}_{:=T_1} \\
&\quad + \frac{1}{1-\gamma}\mathbb{E}_{s'\sim d^{\pi}}\underbrace{\left[\sum_a \eta\pi_t(a|s')\left(\hat{A}^{\pi_t}(s',a) - A^{\pi_t}(s',a)\right)A^{\pi_t}(s',a)\right]}_{:=T_2}
\end{aligned}
$$

T_1 can be bounded using Lemma 4,

$$T_1 \geq \ \eta C_t \max_a A^{\pi_t}(s,a)$$

To bound T_2, we use *Assumption* 7,

$$
\begin{aligned}
T_2 &\geq -\eta\left[\sum_a \pi_t(a|s')\,|(\hat{A}^{\pi_t}(s',a) - A^{\pi_t}(s',a))||A^{\pi_t}(s',a)|\right] \\
&\geq -\eta\left[\sum_a \pi_t(a|s')\,|(\hat{A}^{\pi_t}(s',a) - A^{\pi_t}(s',a))|\frac{1}{1-\gamma}\right] && \text{(since } A^{\pi} \leq 1/(1-\gamma)\text{)} \\
&\geq -\eta\left[\sum_a \pi_t(a|s')\,\frac{\epsilon_{approx}}{1-\gamma}\right] && \text{(using Assumption 7)} \\
&= -\frac{\eta\epsilon_{approx}}{1-\gamma}
\end{aligned}
$$

Using the lower-bound for T_1 and T_2 we have

$$\sum_a \pi(a|s)A^{\pi_t}(s,a) \geq \eta C_t \max_a A^{\pi_t}(s,a) - \frac{\eta\epsilon_{approx}}{1-\gamma} \tag{A.11}$$

Putting everything together we have,

$$V^{\pi}(s) \geq V^{\pi_t}(s) - \frac{\eta\epsilon_{approx}}{(1-\gamma)^2} \tag{A.12}$$

$$\geq V^{\pi_t}(s) - \frac{\epsilon_{approx}}{(1-\gamma)} \qquad \text{(since } \eta \leq 1-\gamma\text{)}$$

$$\implies V^{\pi_t}(s) - V^{\pi}(s) \leq \frac{\epsilon_{approx}}{(1-\gamma)} \tag{A.13}$$

$$\tag{A.14}$$

Let T be the Bellman optimality operator defined as:

$$(Tv)(s) = \max_a \{r(s,a) + \gamma \sum_{s'} \Pr[s'|s,a] v(s')\}$$

Applying the operator at iteration t we obtain:

$$TV^{\pi_t}(s) - V^{\pi_t}(s) = \max_a Q^{\pi_t}(s,a) - V^{\pi_t}(s) = \max_a A^{\pi_t}(s,a) \tag{A.15}$$

Let T^{π} be an operator w.r.t π defined as:

$$T^{\pi}(v) = \sum_a \pi(a|s) r(s,a) + \gamma \sum_a \pi(a|s) \sum_{s'} \Pr[s'|s,a] v(s')$$

Applying T^{π} to $V^{\pi'}(s)$ results in:

$$\begin{aligned} T^{\pi}V^{\pi'}(s) &= \sum_a \pi(a|s) r(s,a) + \gamma \sum_a \pi(a|s) \sum_{s'} \Pr[s'|s,a] V^{\pi'}(s) \\ &= \sum_a \pi(a|s) Q^{\pi'}(s,a) \end{aligned}$$

Using the above we obtain:

$$\begin{aligned} T^{\pi}V^{\pi_t}(s) - V^{\pi_t}(s) &= \sum_a \pi(a|s) A^{\pi_t}(s,a) \\ &\geq \eta\, C_t \max_a A^{\pi_t}(s,a) - \frac{\eta\epsilon_{approx}}{1-\gamma} && \text{(using Equation (A.11))} \\ &= \eta\, C_t\, [TV^{\pi_t}(s) - V^{\pi_t}(s)] - \frac{\eta\epsilon_{approx}}{1-\gamma} && \text{(using Eq. A.15)} \\ &\geq \eta\, C_t\, [TV^{\pi_t}(s) - V^{\pi_t}(s)] - \epsilon_{approx} && \text{(since } \eta \leq 1-\gamma\text{)} \end{aligned}$$

Using Equation (A.13) we have

$$T^{\pi}V^{\pi_t}(s) - T^{\pi}V^{\pi}(s) = \gamma \sum_a \pi(a|s) \sum_{s'} \Pr[s'|s,a](V^{\pi_t}(s') - V^{\pi}(s')) \tag{A.16}$$

$$\leq \gamma \sum_a \pi(a|s) \sum_{s'} \Pr[s'|s,a] \frac{\epsilon_{approx}}{1-\gamma} \qquad \text{(using Equation (A.13))}$$

$$= \frac{\gamma\epsilon_{approx}}{1-\gamma} \tag{A.17}$$

Assuming π^* is the optimal policy we have:

$$
\begin{aligned}
V^*(s) - V^{\pi}(s) &= V^*(s) - T^{\pi}V^{\pi}(s) && \text{(since } T^{\pi}V^{\pi}(s) = V^{\pi}(s)\text{)} \\
&= V^*(s) - T^{\pi}V^{\pi_t}(s) + T^{\pi}V^{\pi_t}(s) - T^{\pi}V^{\pi}(s) \\
&\leq V^*(s) - T^{\pi}V^{\pi_t}(s) + \frac{\gamma\epsilon_{approx}}{1-\gamma} && \text{(using Equation (A.17))} \\
&= V^*(s) - V^{\pi_t}(s) - [T^{\pi}V^{\pi_t}(s) - V^{\pi_t}(s)] + \frac{\gamma\epsilon_{approx}}{1-\gamma} \\
& && \text{(add and subtract } V^{\pi_t}(s)\text{)} \\
&\leq V^*(s) - V^{\pi_t}(s) - \eta\, C_t\, [TV^{\pi_t}(s) - V^{\pi_t}(s)] + \epsilon_{approx} + \frac{\gamma\epsilon_{approx}}{1-\gamma} \\
&= \eta\, C_t[V^*(s) - V^{\pi_t}(s)] + (1 - \eta\, C_t)[V^*(s) - V^{\pi_t}(s)] - \eta\, C_t\, [TV^{\pi_t}(s) - V^{\pi_t}(s)] + \frac{\epsilon_{approx}}{1-\gamma} \\
&= \eta\, C_t\, [TV^*(s) - V^{\pi_t}(s) - TV^{\pi_t}(s) + V^{\pi_t}(s)] + (1 - \eta\, C_t)\, [V^*(s) - V^{\pi_t}(s)] + \frac{\epsilon_{approx}}{1-\gamma} \\
&= \eta\, C_t\, [TV^*(s) - TV^{\pi_t}(s)] + (1 - \eta\, C_t)\, [V^*(s) - V^{\pi_t}(s)] + \frac{\epsilon_{approx}}{1-\gamma} \\
&\leq \gamma\, \eta\, C_t\, [V^*(s) - V^{\pi_t}(s)] + (1 - \eta\, C_t)\, [V^*(s) - V^{\pi_t}(s)] + \frac{\epsilon_{approx}}{1-\gamma} \\
& && (T \text{ is a } \gamma \text{ contraction map}) \\
&= [1 - \eta\, C_t(1-\gamma)]\, [V^*(s) - V^{\pi_t}(s)] + \frac{\epsilon_{approx}}{1-\gamma}
\end{aligned}
$$

□

Theorem 10. *Under Assumption 1-7, Algorithm 1 with* $\eta < \min\left\{1-\gamma, \frac{1}{C_t(1-\gamma)}\right\}$ *converges as,*

$$
\begin{aligned}
& J(\pi^*) - J(\pi_T) \\
& \leq \left(\prod_{t=0}^{T-1} \alpha_t\right) (J(\pi^*) - J(\pi_0)) + \beta \sum_{t=0}^{T-1} \prod_{i=t+1}^{T-1} \alpha_t\,,
\end{aligned}
$$

where $\beta = \frac{\sqrt{2}}{(1-\gamma)^2\rho_{min}}\sqrt{\epsilon_{stat} + \epsilon_{bias}} + \frac{\epsilon_{approx}}{1-\gamma}$, $\alpha_t = [1 - \eta\, C_t(1-\gamma)]$, $C_t := \min_s\{\pi_t(\tilde{\mathcal{A}}_t(s)|s)\, \Delta^t(s)\}$, $\tilde{\mathcal{A}}_t(s) := \arg\max_a Q^{\pi_t}(s,a)$, $\pi_t(\tilde{\mathcal{A}}_t(s)|s) = \sum_{a \in \tilde{\mathcal{A}}_t(s)} \pi_t(a_t(s)|s)$ *and* $\Delta^t(s) := \max_a Q^{\pi_t}(s,a) - \max_{a \notin \tilde{\mathcal{A}}} Q^{\pi_t}(s,a)$.

Proof. Using Lemma 6 with $\pi = \pi_{t+1/2}$,

$$
V^*(s) - V^{\pi_{t+1/2}}(s) \leq [1 - \eta\, C_t(1-\gamma)]\, [V^*(s) - V^{\pi_t}(s)] + \frac{\epsilon_{approx}}{1-\gamma}
$$

The rest of the proof is similar to the proof of Theorem 3. For completeness, we repeat it here. By adding and removing $V^{\pi_{t+1}}(s)$ to both sides and rearranging we have

$$
V^*(s) - V^{\pi_{t+1}}(s) \leq [1 - \eta\, C_t(1-\gamma)]\, [V^*(s) - V^{\pi_t}(s)] + V^{\pi_{t+1/2}}(s) - V^{\pi_{t+1}}(s) + \frac{\epsilon_{approx}}{1-\gamma}.
$$

Taking the expectation w.r.t. ρ we obtain:

$$J(\pi^*) - J(\pi_{t+1}) \leq [1 - \eta\ C_t(1-\gamma)]\,[J(\pi^*) - J(\pi_t)] + \underbrace{J(\pi_{t+1/2}) - J(\pi_{t+1})}_{:=E_1} + \frac{\epsilon_{approx}}{1-\gamma}.$$

The term E_1 can be bounded as follows:

$$\begin{aligned}
E_1 &= \sum_s d^{\pi_{t+1/2}}(s) \langle Q^{\pi_{t+1}}(s,.), \pi_{t+1/2}(.|s) - \pi_{t+1}(.|s)\rangle \\
&\leq \sum_s d^{\pi_{t+1/2}}(s)\, \|Q^{\pi_{t+1}}(s,.)\|_\infty \left\|\pi_{t+1/2}(.|s) - \pi_{t+1}(.|s)\right\|_1 && \text{(Holder inequality)} \\
&\leq \frac{1}{1-\gamma} \sum_s d^{\pi_{t+1/2}}(s) \left\|\pi_{t+1/2}(.|s) - \pi_{t+1}(.|s)\right\|_1 && \text{(since } \|Q^{\pi_{t+1}}(s,.)\|_\infty \leq \tfrac{1}{1-\gamma}\text{)} \\
&= \frac{1}{1-\gamma} \sum_s \frac{d^{\pi_{t+1/2}}(s)}{\rho(s)} \rho(s) \left\|\pi_{t+1/2}(.|s) - \pi_{t+1}(.|s)\right\|_1 \\
&\leq \frac{1}{1-\gamma} \left\|\frac{d^{\pi_{t+1/2}}}{\rho}\right\|_\infty \sum_s \rho(s) \left\|\pi_{t+1/2}(.|s) - \pi_{t+1}(.|s)\right\|_1 \\
&\leq \frac{1}{(1-\gamma)\rho_{min}} \sum_s \rho(s) \left\|\pi_{t+1/2}(.|s) - \pi_{t+1}(.|s)\right\|_1 \\
& && \text{(since } d^{\pi_{t+1/2}}(s) \leq 1 \text{ and using Assumption 3)} \\
&\leq \frac{1}{(1-\gamma)^2\rho_{min}} \sum_s d^{\pi_t}(s) \left\|\pi_{t+1/2}(.|s) - \pi_{t+1}(.|s)\right\|_1 && \text{(sicne } d^{\pi_t} \geq (1-\gamma)\rho\text{)} \\
&\leq \frac{\sqrt{2}}{(1-\gamma)^2\rho_{min}} \sum_s d^{\pi_t}(s) \sqrt{\mathrm{KL}(\pi_{t+1/2}(.|s)\,||\,\pi_{t+1}(.|s))} \\
& && \text{(using strong convexity of KL divergence or Pinsker's inequality)} \\
&\leq \frac{\sqrt{2}}{(1-\gamma)^2\rho_{min}} \sqrt{\underbrace{\sum_s d^{\pi_t}(s) \mathrm{KL}(\pi_{t+1/2}(.|s)\,||\,\pi_{t+1}(.|s))}_{:=E_2}} && \text{(due to concavity of } \sqrt{}\text{)}
\end{aligned}$$

where E_2 can be bounded as follows:

$$\begin{aligned}
E_2 &= \tilde{\ell}_t(\theta_{t+1}) \\
&= \tilde{\ell}_t(\theta_{t+1}) - \min_\theta \tilde{\ell}_t(\theta_{t+1}) + \min_\theta \tilde{\ell}_t(\theta_{t+1}) \\
&\leq \epsilon_{stat} + \min_\theta \tilde{\ell}_t(\theta_{t+1}) && \text{(using Assumption 1)} \\
&\leq \epsilon_{stat} + \epsilon_{bias} && \text{(using Assumption 2)}
\end{aligned}$$

Putting everything together we have:

$$E_1 \leq \underbrace{\frac{\sqrt{2}}{(1-\gamma)^2\rho_{min}} \sqrt{\epsilon_{\text{stat}} + \epsilon_{\text{bias}}}}_{:=\beta'}$$

Therefore we have

$$J(\pi^*) - J(\pi_{t+1}) \leq \underbrace{[1 - \eta\ C_t(1-\gamma)]}_{\alpha_t} [J(\pi^*) - J(\pi_t)] + \underbrace{\beta' + \frac{\epsilon_{approx}}{1-\gamma}}_{:=\beta}.$$

Unrolling the above recursion for T iterations,

$$J(\pi^*) - J(\pi_T) \leq \left(\prod_{t=0}^{T-1} \alpha_t \right) (J(\pi^*) - J(\pi_0)) + \beta \sum_{t=0}^{T-1} \prod_{i=t+1}^{T-1} \alpha_i$$

□

A.3 Tabular MDP Experiments

In this section, we empirically evaluate SPMA on tabular MDP environments. For these experiments, we use Cliff World [Sutton and Barto, 2018] and Frozen Lake [Brockman, 2016] following the setup in Vaswani et al. [2024]. In subsection, A.3.1 we examine the case where the policy is parametrized using softmax tabular representation. In subsection, A.3.2 we investigate the function approximation setting as described in Section 2.4, where the policy is parametrized using a linear model.

A.3.1 Softmax Tabular Representation

For this parametrization we initialize $z \in \mathbb{R}^{S \times A}$ uniformly, i.e., $\pi_0(a|s) = \frac{1}{|\mathcal{A}|}$ for each a and s. Furthermore, for each algorithm, we set η using a grid search and pick the step-sizes that result in the best area under the curve (AUC). The tabular MDP results suggest that SPMA and NPG achieve similar performance and both outperform SPG [Sutton et al., 1999, Schulman et al., 2017b] (see Fig. A.1).

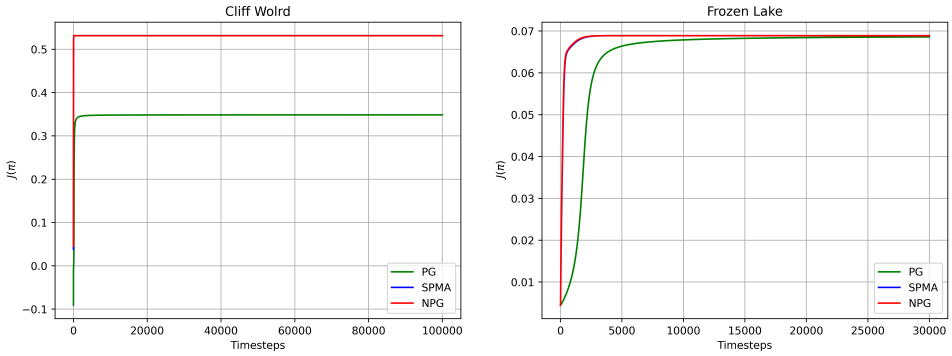


Figure A.1: SPMA matches the performance of NPG and they both outperform SPG.

Sensitivity to the step size. Figure A.2 compares the algorithms across choices of η. The greater sensitivity of SPG (green), relative to SPMA (blue) and NPG (red), complements the convergence comparison by showing that the mirror-ascent updates in Chapter 2 remain effective over a broader range of step sizes.

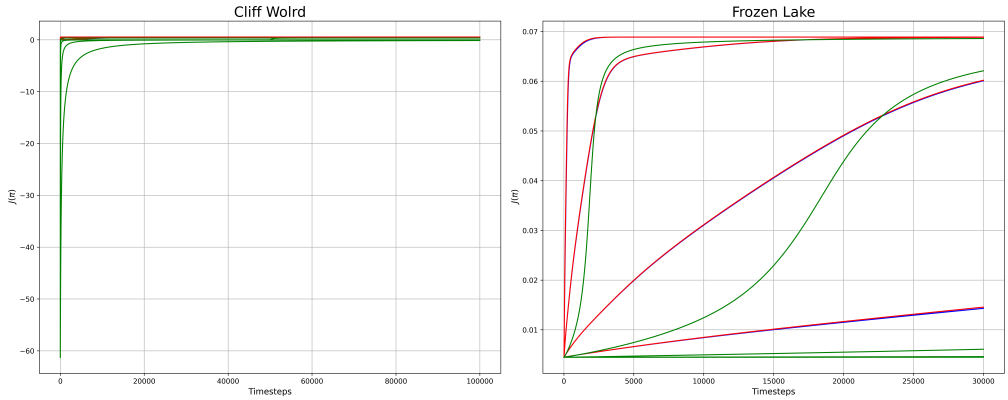


Figure A.2: SPG (green) is more sensitive to η compared to SPMA and NPG (blue and red).

A.3.2 Linear Functional Approximation (Linear FA)

For the Linear FA setting, we use a log-linear policy parametrization: $\pi_t(a|s) = \frac{\exp(\mathbf{X}(s,a)\theta)}{\sum_{a'}\exp(\mathbf{X}(s,a')\theta)}$, with $\mathbf{X} \in \mathbb{R}^{\mathcal{SA}\times d}$ and $\theta \in \mathbb{R}^d$ representing the features and parameters. We use constant initialization for θ and following Vaswani et al. [2024], use tile-coded features for $\mathbf{X}$. As in the previous section, we set η for SPMA and MDPO via grid search and report results based on the best AUC. For the inner loop optimization (e.g., minimizing Equation (2.5)), we use Armijo line search [Armijo, 1966], avoiding an additional grid search for the step-size. For SPG we use the update from Mei et al. [2020] where Armijo line search is used to set η.

We make the following observations from the results: (i) SPG performs poorly in the linear FA setting, while both SPMA and MDPO perform well when the parameter dimension d and the number of inner loop optimizations m are sufficiently large. (ii) In the CW environment, for smaller d, SPMA converges faster than MDPO (Fig. A.3, top row). Increasing m from 25 to 50 narrows the gap between SPMA and MDPO (top vs. bottom row). (iii) In the FL environment, SPMA and MDPO perform similarly, both outperforming SPG (Fig. A.4).

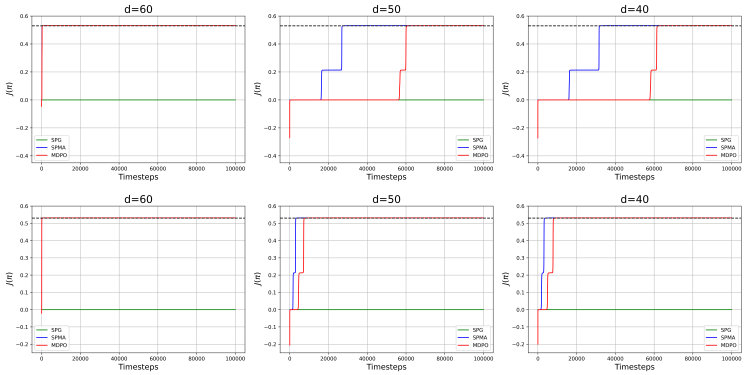


Figure A.3: CW environment: The top row ($m = 25$) shows that SPMA converges faster than MDPO as d decreases, while the bottom row ($m = 50$) shows the gap decreases as the number of inner loop optimizations increases.

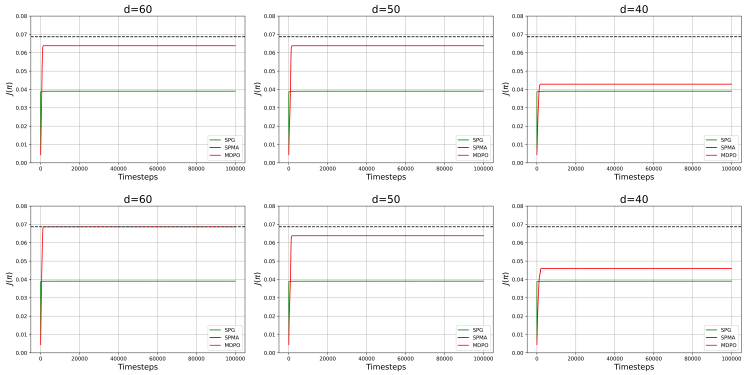


Figure A.4: FL environment: The top row ($m = 25$) and bottom row ($m = 50$) show that SPMA and MDPO have similar convergence and both outperform SPG. The performance of both SPMA and MDPO improves as d increases (i.e., the bias decreases) and m increases (i.e., the optimization error decreases).

A.3.3 Empirical Verification that C_t is Non-zero

In this section, we empirically investigate the value of C_t. Recall that $C_t := \min_s\{\pi_t(\tilde{\mathcal{A}}_t(s)|s)\,\Delta^t(s)\}$, $\tilde{\mathcal{A}}_t(s) := \arg\max_a Q^{\pi_t}(s,a)$, $\pi_t(\tilde{\mathcal{A}}_t(s)|s) = \sum_{a\in\tilde{\mathcal{A}}_t(s)}\pi_t(a|s)$, and $\Delta^t(s) := \max_a Q^{\pi_t}(s,a) - \max_{a\notin\tilde{\mathcal{A}}_t(s)} Q^{\pi_t}(s,a)$. To this end, we compute the per-state metric $C_t(s) := \pi_t(\tilde{\mathcal{A}}_t(s)|s)\,\Delta^t(s)$ and its components, $\pi_t(\tilde{\mathcal{A}}_t(s)|s)$ and $\Delta^t(s)$, for the CliffWorld and FrozenLake environments using softmax tabular policy parametrization. Our results indicate that $C_t(s)$ is nonzero for all states except for certain terminal states, specifically the goal state and the hole states in FrozenLake, where $Q^\pi(s,a)$ is zero for all actions a and policies π. In these cases, we observe that $\Delta^t(s)$ becomes zero, leading to $C_t(s) = 0$ and consequently $C_t = 0$. Note that in the CliffWorld environment, the transition probability matrix moves the agent back to the starting point at terminal states, so they do not pose an issue. However, in the FrozenLake environment, the agent remains at the terminal state with probability one.

Given these empirical results, one could modify the theorems in this paper to exclude the problematic terminal states. For example, in Theorem 2, instead of bounding $V^{\pi^*}(s) - V^{\pi_T}(s)$ for all states using the ℓ_∞ norm, we could exclude states where $Q^{\pi^*}(s,a) = 0$ for all actions a. This exclusion is analogous to the bandit literature, which considers only arms with a nonzero gap in the regret bound. In Figure A.5, we plot C_t using softmax tabular policy parameterizations for both CliffWorld and FrozenLake environments, confirming that after excluding terminal states, C_t is lower-bounded by a positive constant.

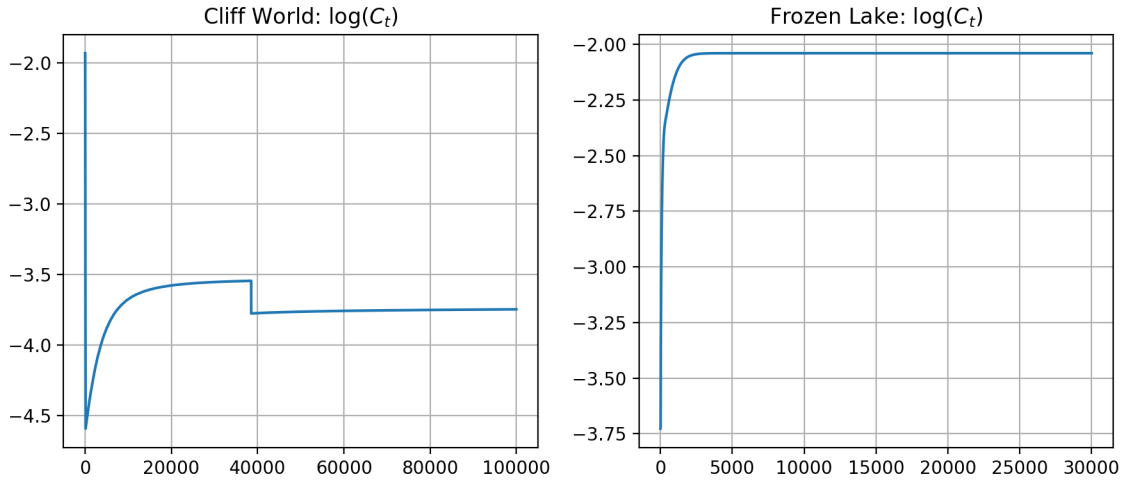


Figure A.5: SPMA with softmax tabular policy parametrization: After excluding terminal states, results confirm that C_t is lower-bounded by a positive constant.

A.4 Additional Details for Stable Baselines Experiments

In subsection A.4.1, we provide additional details on the hyper-parameters used for the results in Section 2.5. Next, we present an ablation study on the number of inner loop optimization steps (m) in subsection A.4.2.

A.4.1 Atari and Mujoco Details

In the Atari experiments, we use the default hyper-parameters for each method from stable baselines [Raffin et al., 2021]. This choice is motivated by two factors: i) following the work of Tomar et al. [2020], we aim to evaluate the effectiveness of different surrogate losses without conducting an exhaustive search over numerous hyper-parameters; ii) the CNN-based actor and critic networks make grid searching over many hyper-parameters (e.g., framestack, GAE λ, horizon length, discount factor) computationally infeasible. For a complete list of hyper-parameters used in the Atari experiments, see Table A.1.

In the MuJoCo experiments, we use the default hyper-parameters from stable baselines for each method, but perform a grid search on the Adam inner loop step size for `PPO` and `TRPO-constrained` (best among $[3 \times 10^{-3}, 3 \times 10^{-4}, 3 \times 10^{-5}]$) and the probability ratio clipping parameter in `PPO` (best from $[0.1, 0.2, 0.3]$). For the regularized surrogates (i.e., the remaining methods: `SPMA`, `MDPO`, and `TRPO-regularized`), we avoid a grid search for the inner loop step size by using a full batch (i.e., the horizon length) along with the Armijo line search [Armijo, 1966]. See Table A.2 for the full list of hyper-parameters used in the MuJoCo experiments.

To set η for the regularized surrogates, we perform a grid search over five fixed values ($[0.3, 0.5, 0.7, 0.9, 1.0]$). Although Tomar et al. [2020] anneals η from 1 to 0 during training, we observe that using a constant step size results in better performance. Our grid search strategy for all stable baselines experiments is consistent: we run the experiments for 2 million iterations, select the hyper-parameters that yield the best AUC, and then use these hyper-parameters for an additional 8 million iterations.

Hyperparameter	**SPMA**	**MDPO**	**TRPO_regularized**	**TRPO_constrained**	**PPO**
Reward normalization	✗	✗	✗	✗	✗
Observation normalization	✗	✗	✗	✗	✗
Orthogonal weight initialization	✓	✓	✓	✓	✓
Value function clipping	✗	✗	✗	✗	✗
Gradient clipping	✗	✗	✗	✗	✓
Probability ratio clipping	✗	✗	✗	✗	✓
Adam step-size	3×10^{-4}				
Minibatch size	256				
Framestack	4				
Number of environment copies	8				
GAE λ	0.95				
Horizon (T)	128				
Number of inner loop updates (m)	5				
Entropy coefficient	0				
Discount factor	0.99				
Total number of timesteps	10^7				
Number of runs for plot averages	5				
Confidence interval for plot runs	$\sim 95\%$				

Table A.1: Hyperparameters for Atari experiments.

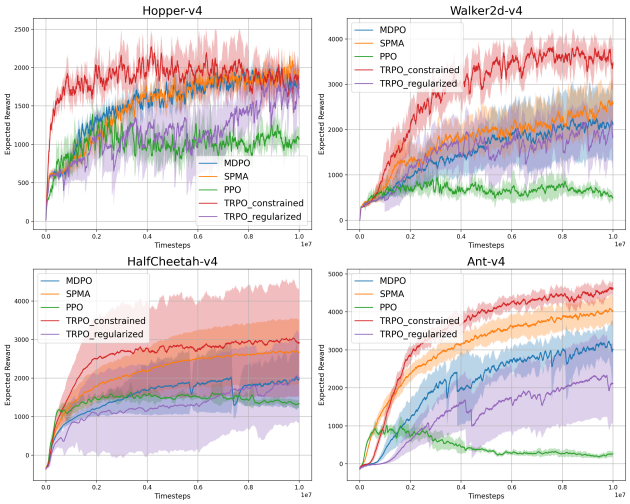
Hyperparameter	**SPMA**	**MDPO**	**TRPO_regularized**	**TRPO_constrained**	**PPO**
Minibatch size	2048	2048	2048	64	64
Reward normalization	✗	✗	✗	✗	✗
Observation normalization	✗	✗	✗	✗	✗
Orthogonal weight initialization	✓	✓	✓	✓	✓
Value function clipping	✗	✗	✗	✗	✗
Gradient clipping	✗	✗	✗	✗	✓
Probability ratio clipping	✗	✗	✗	✗	✓
Adam step-size	✗	✗	✗	✓	✓
GAE λ	0.95				
Horizon (T)	2048				
Number of inner loop updates (m)	5				
Entropy coefficient	0				
Discount factor	0.99				
Total number of timesteps	10^7				
Number of runs for plot averages	5				
Confidence interval for plot runs	$\sim 95\%$				

Table A.2: Hyperparameters for MuJoCo experiments.

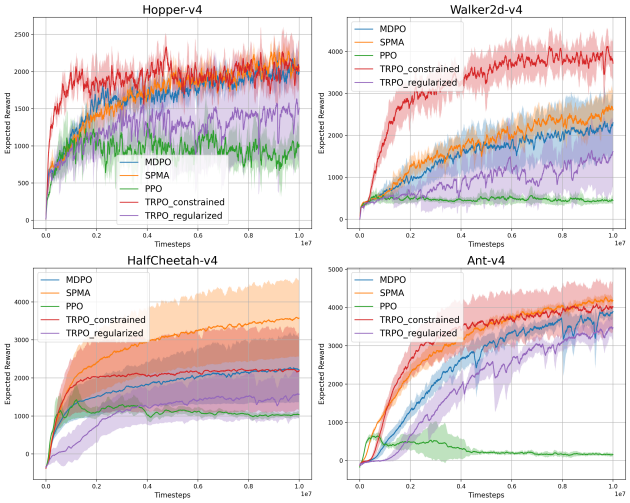
A.4.2 Ablation Study on the Number of Inner Loop Optimization Steps

This subsection studies how the number of inner-loop optimization steps m affects the stable-baselines experiments. Consistent with Tomar et al. [2020], $m = 1$ performs poorly, so the experiments focus on larger optimization budgets.

MuJoCo at larger inner-loop budgets. Figure A.6 extends the main-paper MuJoCo comparison from $m = 5$ to $m = 10$ and $m = 15$ inner-loop optimization steps. As the inner-loop is optimized more accurately, `SPMA` can approach the fine-tuned `TRPO-constrained`.



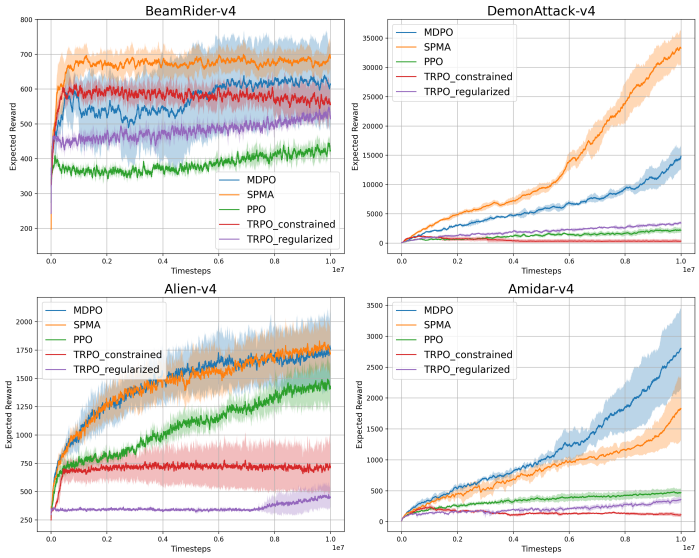
(a)



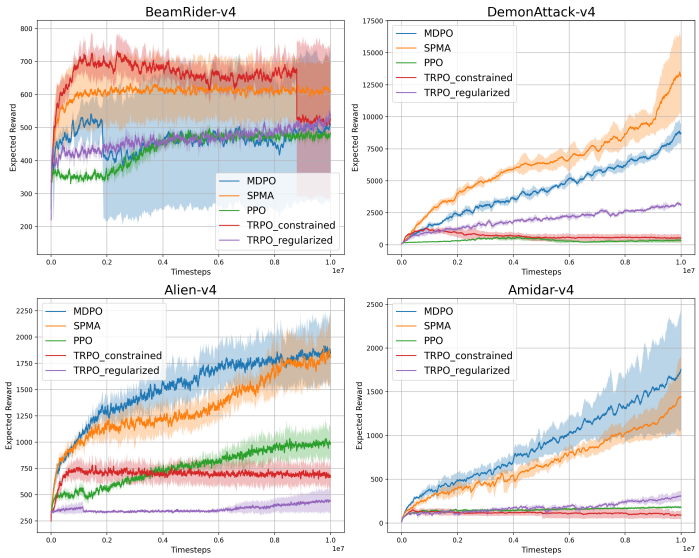
(b)

Figure A.6: MuJoCo results for $m = 10$ (a) and $m = 15$ (b). As m increases from 5 (Figure 2.2) to 10 and 15, `SPMA` shows performance comparable to the fine-tuned `TRPO-constrained`.

Atari at larger inner-loop budgets. Figure A.7 shows that the MuJoCo trend in the previous figure does not transfer uniformly to Atari: increasing the number of inner-loop optimization steps m from 5 to 10 or 15 does not consistently improve performance.



(a)



(b)

Figure A.7: Atari results for $m = 10$ (top) and $m = 15$ (bottom). Increasing m does not necessarily lead to performance improvements.

MuJoCo ablation across optimization budgets. Figure A.8 isolates the effect of m across the four MuJoCo environments. The results indicate consistent gains for SPMA as m increases.

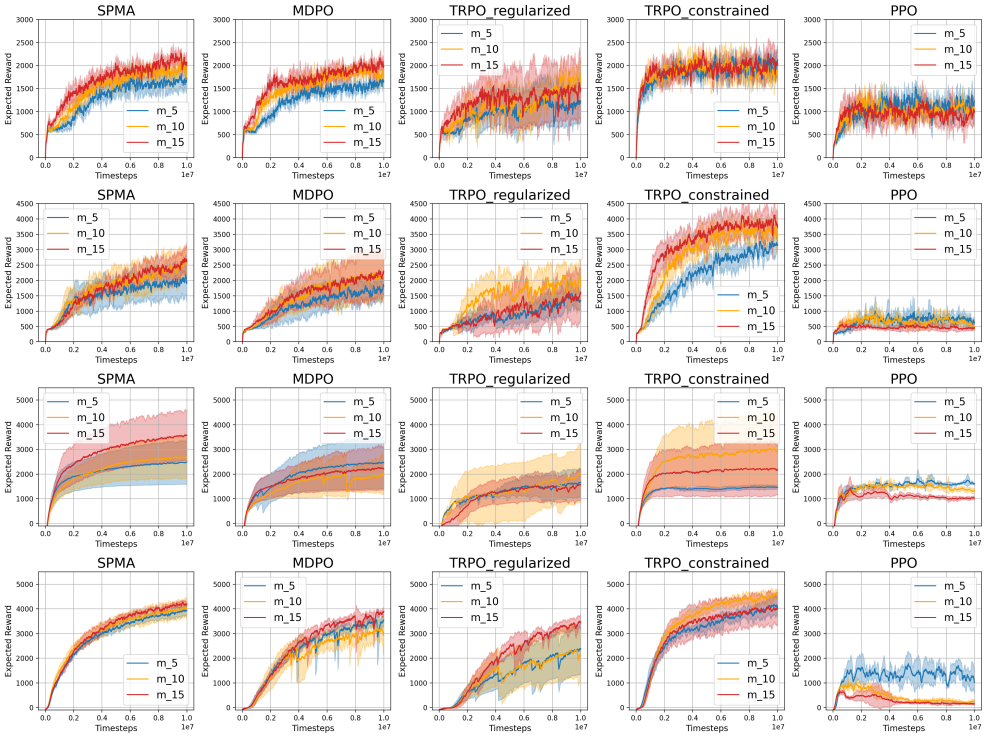


Figure A.8: MuJoCo ablation on m: The rows correspond to the Hopper-v4, Walker2d-v4, HalfCheetah-v4, and Ant-v4 environments, respectively. As the number of inner loop optimization steps m increases, SPMA shows improvements in expected reward and becomes comparable to the fine-tuned TRPO-constrained.

Atari ablation across optimization budgets. Figure A.9 reports the corresponding ablation on Atari. Increasing the number of inner-loop optimization steps does not necessarily lead to a monotonic improvement in performance.

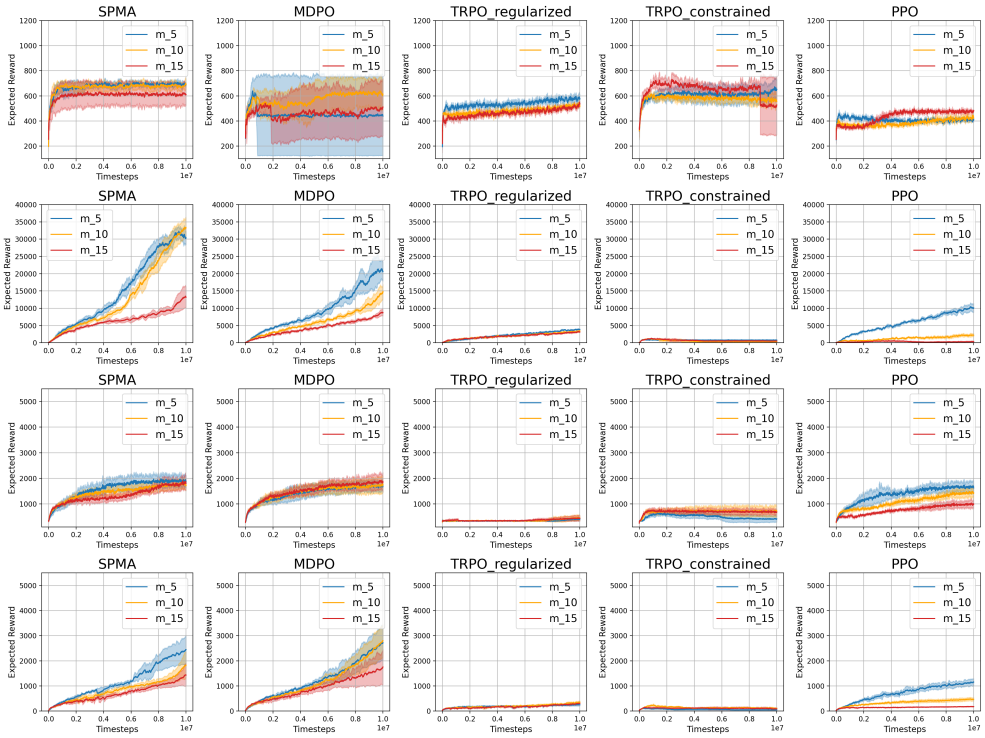


Figure A.9: Atari ablation on m: The rows correspond to the BeamRider-v4, DemonAttack-v4, Alien-v4, and Amidar-v4 games. We observe that increasing m does not necessarily improve results across methods.

Appendix B

Dissecting Discrete Soft Actor-Critic: Limitations and Principled Alternatives

B.1 Actor Objective Instantiations

In this section, we instantiate our forward and reverse KL-based objectives in Equations 3.5a and 3.5b using soft NPG, and soft SPMA in the function approximation setting.

B.1.1 NPG-FKL(τ, ζ)

$$
\begin{aligned}
\pi_{t+1} &= \arg\min_{\pi_\theta \in \Pi} \mathbb{E}_{s\sim\mathcal{D}_t} \Big[\mathrm{KL}(\pi_{t+1/2}(\cdot|s)||\pi_\theta(\cdot|s)) - \tau_t \mathcal{H}(\pi_\theta(\cdot|s))\Big] \\
&= \arg\min_{\pi_\theta \in \Pi} \mathbb{E}_{s\sim\mathcal{D}_t} \left[\mathbb{E}_{a\sim\pi_t(\cdot|s)} \left[-\frac{\exp(\eta_t\, q_\zeta^t(s,a))}{\mathcal{Z}_t(s)} \ln\left(\frac{\pi_\theta(a|s)}{\frac{\pi_t(a|s)\,\exp(\eta_t q_\zeta^t(s,a))}{\mathcal{Z}_t(s)}}\right)\right] - \tau_t \mathcal{H}(\pi_\theta(\cdot|s))\right] \\
&\qquad\qquad \text{(Using the NPG update in Equation (3.3))} \\
&= \arg\min_{\pi_\theta \in \Pi} \mathbb{E}_{s\sim\mathcal{D}_t} \left[\mathbb{E}_{a\sim\pi_t(\cdot|s)} \left[-\frac{\exp(\eta_t\, q_\zeta^t(s,a))}{\mathcal{Z}_t(s)} \ln\left(\frac{\pi_\theta(a|s)}{\pi_t(a|s)}\right)\right] - \tau_t \mathcal{H}(\pi_\theta(\cdot|s))\right] \\
&\qquad\qquad \text{(dropping the constant w.r.t } \pi\text{)} \\
&= \arg\max_{\pi_\theta \in \Pi} \mathbb{E}_{s\sim\mathcal{D}_t} \left[\mathbb{E}_{a\sim\pi_t(\cdot|s)} \left[\frac{\exp(\eta_t\, q_\zeta^t(s,a))}{\sum_{a'} \pi_t(a')\,\exp(\eta_t\, q_\zeta^t(s,a))} \ln\left(\frac{\pi_\theta(a|s)}{\pi_t(a|s)}\right)\right] + \tau_t \mathcal{H}(\pi_\theta(\cdot|s))\right]
\end{aligned}
$$

B.1.2 SPMA-FKL(τ, ζ)

$$
\begin{aligned}
\pi_{t+1} &= \arg\min_{\pi_\theta \in \Pi} \mathbb{E}_{s\sim\mathcal{D}_t} \Big[\mathrm{KL}(\pi_{t+1/2}(\cdot|s)||\pi_\theta(\cdot|s)) - \tau_t \mathcal{H}(\pi_\theta(\cdot|s))\Big] \\
&= \arg\min_{\pi_\theta \in \Pi} \mathbb{E}_{s\sim\mathcal{D}_t} \left[\mathbb{E}_{a\sim\pi_t(\cdot|s)} \left[- \frac{\left(1+\eta_t(q^t_\zeta(s,a) - v_\zeta(s)^t)\right)}{\mathcal{Z}_t(s)} \ln\left(\frac{\frac{\pi_\theta(a|s)}{\pi_t(a|s)\,(1+\eta_t\,(q^t_\zeta(s,a) - v^t_\zeta(s)))}}{\frac{}{\mathcal{Z}_t(s)}}\right)\right]\right. \\
&\qquad\qquad \left. - \tau_t \mathcal{H}(\pi_\theta(\cdot|s))\right] \\
&\qquad\qquad \text{(Using the SPMA update in Equation (3.4))} \\
&= \arg\min_{\pi_\theta \in \Pi} \mathbb{E}_{s\sim\mathcal{D}_t} \left[\mathbb{E}_{a\sim\pi_t(\cdot|s)} \left[- \frac{\Big(1 + \eta_t(q^t_\zeta(s,a) - v^t_\zeta(s))\Big)}{\mathcal{Z}_t(s)} \ln\left(\frac{\pi_\theta(a|s)}{\pi_t(a|s)}\right)\right]\right. \\
&\qquad\qquad \left. - \tau_t \mathcal{H}(\pi_\theta(\cdot|s))\right] \qquad \text{(dropping the constant w.r.t } \pi\text{)} \\
&= \arg\max_{\pi_\theta \in \Pi} \mathbb{E}_{s\sim\mathcal{D}_t} \left[\mathbb{E}_{a\sim\pi_t(\cdot|s)} \left[\frac{\Big(1 + \eta_t(q^t_\zeta(s,a) - v^t_\zeta(s))\Big)}{\sum_{a'}\ \pi_t(a'|s)\ \Big(1 + \eta_t(q^t_\zeta(s,a') - v^t_\zeta(s))\Big)} \ln\left(\frac{\pi(a|s)}{\pi_t(a|s)}\right)\right]\right. \\
&\qquad\qquad \left. + \tau_t \mathcal{H}(\pi_\theta(\cdot|s))\right]
\end{aligned}
$$

B.1.3 NPG-RKL(τ, ζ)

$$
\begin{aligned}
\pi_{t+1} &= \arg\min_{\pi_\theta \in \Pi} \mathbb{E}_{s\sim\mathcal{D}_t} \Big[\mathrm{KL}(\pi_\theta(\cdot|s)||\pi_{t+1/2}(\cdot|s)) - \tau_t \mathcal{H}(\pi_\theta(\cdot|s))\Big] \\
&= \arg\min_{\pi_\theta \in \Pi} \mathbb{E}_{s\sim\mathcal{D}_t}\, \mathbb{E}_{a\sim\pi_\theta(.|s)} \Big[(1+\tau_t)\ \ln(\pi_\theta(a|s)) - \ln(\pi_{t+1/2}(a|s))\Big] \\
&= \arg\min_{\pi_\theta \in \Pi} \mathbb{E}_{s\sim\mathcal{D}_t} \mathbb{E}_{a\sim\pi_\theta(.|s)} \left[(1+\tau_t)\ \ln(\pi_\theta(a|s)) - \ln\left(\pi_t(a|s)\exp(\eta_t\, q^t_\zeta(s,a))\right)\right] \\
&\qquad \text{(Using the NPG update in Equation (3.3) and since } \mathcal{Z}_t \text{ can be marginalized out)} \\
&= \arg\min_{\pi_\theta \in \Pi} \mathbb{E}_{s\sim\mathcal{D}_t} \mathbb{E}_{a\sim\pi_\theta(.|s)} \Big[(1+\tau_t)\ \ln(\pi_\theta(a|s)) - \ln(\pi_t(a|s)) - \eta_t\, q^t_\zeta(s,a))\Big] \\
&= \arg\min_{\pi_\theta \in \Pi} \mathbb{E}_{s\sim\mathcal{D}_t} \mathbb{E}_{a\sim\pi_\theta(.|s)} \left[(1+\tau\eta_t)\ \ln(\pi_\theta(a|s)) - \eta_t\ \left(q^t_\zeta(s,a) + \frac{1}{\eta_t}\ln(\pi_t(a|s))\right)\right] \\
&\qquad\qquad \text{(Since } \tau_t = \eta_t\,\tau\text{)} \\
&= \arg\max_{\pi_\theta \in \Pi} \mathbb{E}_{s\sim\mathcal{D}_t} \mathbb{E}_{a\sim\pi_\theta(.|s)} \left[q^t_\zeta(s,a) - \tau\ \ln(\pi_\theta(a|s)) - \frac{1}{\eta_t}\ln\left(\frac{\pi_\theta(a|s)}{\pi_t(a|s)}\right)\right] \\
&= \arg\max_{\pi_\theta \in \Pi} \mathbb{E}_{s\sim\mathcal{D}_t} \left[\mathbb{E}_{a\sim\pi_\theta(.|s)} \Big[q^t_\zeta(s,a) - \tau\ \ln(\pi_\theta(a|s))\Big] - \frac{1}{\eta_t}\,\mathrm{KL}(\pi_\theta(\cdot|s)\,||\,\pi_t(\cdot|s))\right]
\end{aligned}
$$

Given an estimate of the entropy-regularized q function, DSAC is a special of NPG-RKL by setting $\eta_t = \infty$ resulting in the following surrogate loss:

$$\pi_{t+1} = \arg\max_{\pi_\theta \in \Pi} \mathbb{E}_{s\sim\mathcal{D}_t} \mathbb{E}_{a\sim\pi_\theta(.|s)} \left[\left(q_\zeta^t(s,a) \right) - \tau \, \ln(\pi_\theta(a|s)) \right]$$

B.1.4 SPMA-RKL(τ, ζ)

$$\begin{aligned}
\pi_{t+1} &= \arg\min_{\pi_\theta \in \Pi} \mathbb{E}_{s\sim\mathcal{D}_t} \left[\mathrm{KL}(\pi_\theta(\cdot|s)||\pi_{t+1/2}(\cdot|s)) - \tau_t \mathcal{H}(\pi_\theta(\cdot|s)) \right] \\
&= \arg\min_{\pi_\theta \in \Pi} \mathbb{E}_{s\sim\mathcal{D}_t} \, \mathbb{E}_{a\sim\pi_\theta(.|s)} \left[(1+\tau_t) \, \ln(\pi(a|s)) - \ln(\pi_{t+1/2}(a|s)) \right] \\
&= \arg\min_{\pi_\theta \in \Pi} \mathbb{E}_{s\sim\mathcal{D}_t} \mathbb{E}_{a\sim\pi_\theta(.|s)} \left[(1+\tau_t) \, \ln(\pi_\theta(a|s)) - \ln \left(\pi_t(a|s) \left[1 + \eta_t \left(q_\zeta^t(s,a') - v_\zeta^t(s) \right) \right] \right) \right] \\
&\qquad\qquad \text{(Using the SPMA update in Equation (3.4))} \\
&= \arg\max_{\pi_\theta \in \Pi} \mathbb{E}_{s\sim\mathcal{D}_t} \left[\mathbb{E}_{a\sim\pi_\theta(\cdot|s)} \ln \left(1 + \eta_t (q_\zeta^t(s,a) - v_\zeta^t(s)) \right) - \mathrm{KL}(\pi_\theta||\pi_t) - \tau_t \, \ln \pi_\theta(a|s) \right]
\end{aligned}$$

B.2 General Off-Policy Actor-Critic Pseudocode

Algorithm 5: General Off-Policy Actor-Critic Framework

Input: θ_0 (π_0's parameters), ϕ_0 (q_ζ^0's parameters), π_θ (function approximation for actor), q_ϕ (function approximation for critic), $\mathcal{L}_t$ (critic loss), ℓ_t (actor loss), K (total iterations), N (number of environment steps), n (number of policy optimization steps), α (inner-loop step-size)
for $t = 0$ **to** $K-1$ **do**
  Interact with the environment for N steps to collect data using π_t: $D_t \leftarrow D_t \cup \{s_i, a_i, r(s_i,a_i), s_{i+1}\}_{i=1}^N$
  $\phi_t = \arg\min \mathcal{L}_t(\phi)$; $q_\zeta^t = q_{\phi_t}$
  Initialize inner-loop: $\omega_0 = \theta_t$
  for $j = 0$ **to** $n-1$ **do**
    $\omega_{j+1} = \omega_j - \alpha \nabla_\omega \ell_t(\omega_j)$
  end for
  $\theta_{t+1} = \omega_n$
  $\pi_{t+1}(\cdot|s) = \pi_{\theta_{t+1}}(\cdot|s)$
end for
Return: θ_K

B.3 Theoretical Results

B.3.1 Proof of Theorem 4

Theorem 4 (Generic Reduction with Actor Entropy)**.** *If π_τ^* is the optimal entropy-regularized policy whose value function is v_τ^*, for a q_ζ^t obtained via the policy evaluation scheme at iteration t, define $\epsilon_t := q_\zeta^t - q_\tau^{\pi_t}$. For any sequence of policies $\{\pi_0, \pi_1, \ldots, \pi_{K-1}\}$, if $\bar{\pi}_K$ is the corresponding uniform mixture policy, then,*

$$\left\|v_\tau^* - v_\tau^{\bar{\pi}_K}\right\|_\infty \leq \frac{\|Regret(K)\|_\infty}{K\,(1-\gamma)} + \frac{2\,\sum_{t\in[K]}\|\boldsymbol{\epsilon}_t\|_\infty}{K\,(1-\gamma)}\,,$$

$$\textit{where, } (Regret(K))(s) := \sum_{t=0}^{K-1}\Big[\langle \pi_\tau^*(\cdot|s) - \pi_t(\cdot|s), q_\zeta^t(s,\cdot)\rangle + \tau\big(\mathcal{H}(\pi_\tau^*(\cdot|s)) - \mathcal{H}(\pi_t(\cdot|s))\big)\Big],$$

is the regret incurred on an online optimization problem for each state $s \in \mathcal{S}$.

Proof.

$$\begin{aligned}
v_\tau^{\pi_\tau^*} - v_\tau^{\pi} &= \mathcal{T}_\tau^{\pi_\tau^*} v_\tau^{\pi_\tau^*} - v_\tau^{\pi} && \text{(Since } v_\tau^{\pi_\tau^*} \text{ is a fixed point of } \mathcal{T}_\tau^{\pi_\tau^*}) \\
&= [\mathcal{T}_\tau^{\pi_\tau^*} v_\tau^{\pi} - v_\tau^{\pi}] + [\mathcal{T}_\tau^{\pi_\tau^*} v_\tau^{\pi^*} - \mathcal{T}_\tau^{\pi_\tau^*} v_\tau^{\pi}] && \text{(Add/subtract } \mathcal{T}_\tau^{\pi_\tau^*} v_\tau^{\pi}) \\
&= [\mathcal{T}_\tau^{\pi_\tau^*} v_\tau^{\pi} - v_\tau^{\pi}] + \gamma \mathcal{P}_{\pi_\tau^*}(v_\tau^{\pi_\tau^*} - v_\tau^{\pi}) && \text{(Using the definition of } \mathcal{T}_\tau^{\pi_\tau^*}) \\
\implies v_\tau^{\pi_\tau^*} - v_\tau^{\pi} &= \big(I - \gamma \mathcal{P}_{\pi_\tau^*}\big)^{-1}\left[\mathcal{T}_\tau^{\pi_\tau^*} v_\tau^{\pi} - \mathcal{T}_\tau^{\pi}\, v_\tau^{\pi}\right]
\end{aligned}$$

Summing up from $t = 0$ to $t = K-1$ and dividing by K,

$$\begin{aligned}
v_\tau^* - \frac{\sum_{t=0}^{K-1} v_\tau^{\pi_t}}{K} &= \frac{1}{K}\,(I - \gamma \mathcal{P}_{\pi_\tau^*})^{-1} \sum_{t=0}^{K-1}\left[\mathcal{T}_\tau^{\pi_\tau^*}\, v_\tau^{\pi_t} - v_\tau^{\pi_t}\right] \\
& && \text{(By definition of } v_\tau^*) \\
\implies v_\tau^* - v_\tau^{\bar{\pi}_K} &= \frac{1}{K}(I - \gamma \mathcal{P}_{\pi_\tau^*})^{-1} \sum_{t=0}^{K-1}\left[\mathcal{T}_\tau^{\pi_\tau^*}\, v_\tau^{\pi_t} - v_\tau^{\pi_t}\right] && \text{(Since } v_\tau^{\bar{\pi}_K} = \tfrac{\sum_{t=0}^{K-1} v_\tau^{\pi_t}}{K}) \\
&= \frac{1}{K}\,(I - \gamma \mathcal{P}_{\pi_\tau^*})^{-1} \sum_{t=0}^{K-1}\left[\mathcal{T}_\tau^{\pi_\tau^*}\, v_\tau^{\pi_t} - \mathcal{T}_\tau^{\pi_t}\, v_\tau^{\pi_t}\right] && \text{(Since } v_\tau^{\pi} = \mathcal{T}_\tau^{\pi} v_\tau^{\pi}) \\
\implies \left\|v_\tau^* - v_\tau^{\bar{\pi}_K}\right\|_\infty &= \frac{1}{K}\left\|(I - \gamma \mathcal{P}_{\pi_\tau^*})^{-1} \sum_{t=0}^{K-1}\left[\mathcal{T}_\tau^{\pi_\tau^*}\, v_\tau^{\pi_t} - \mathcal{T}_\tau^{\pi_t}\, v_\tau^{\pi_t}\right]\right\|_\infty \\
&\leq \frac{1}{K}\left\|(I - \gamma \mathcal{P}_{\pi_\tau^*})^{-1}\right\|_\infty \left\|\sum_{t=0}^{K-1}\left[\mathcal{T}_\tau^{\pi_\tau^*}\, v_\tau^{\pi_t} - \mathcal{T}_\tau^{\pi_t}\, v_\tau^{\pi_t}\right]\right\|_\infty \\
& && \text{(By definition of matrix norm)} \\
&\leq \frac{1}{K\,(1-\gamma)}\left\|\sum_{t=0}^{K-1}\left[\mathcal{T}_\tau^{\pi_\tau^*}\, v_\tau^{\pi_t} - \mathcal{T}_\tau^{\pi_t}\, v_\tau^{\pi_t}\right]\right\|_\infty \\
& && \text{(Since } \left\|(I - \gamma \mathcal{P}_\pi)^{-1}\right\|_\infty = \left\|\textstyle\sum_{t=0}^{\infty}[\gamma\, \mathcal{P}_\pi]^t\right\|_\infty \leq \textstyle\sum_{t=0}^{\infty}\gamma^t = \frac{1}{1-\gamma})
\end{aligned}$$

Let us calculate $\left[\mathcal{T}_\tau^{\pi_\tau^*} v_\tau^{\pi_t} - \mathcal{T}_\tau^{\pi_t} v_\tau^{\pi_t}\right](s)$.

$$\begin{aligned}
\left[\mathcal{T}_\tau^{\pi_\tau^*} v_\tau^{\pi_t} - \mathcal{T}_\tau^{\pi_t} v_\tau^{\pi_t}\right](s) &= (\mathcal{T}_\tau^{\pi_\tau^*} v_\tau^{\pi_t})(s) - (\mathcal{T}_\tau^{\pi_t} v_\tau^{\pi_t})(s) \\
&= \mathbb{E}_{a\sim\pi_\tau^*}\left[q_\tau^{\pi_t}(s,a) - \tau\,\ln(\pi_\tau^*(a|s))\right] \\
&\quad - \mathbb{E}_{a\sim\pi_t}\left[q_\tau^{\pi_t}(s,a) - \tau\,\ln(\pi_t(a|s))\right] \qquad \text{(By definition of } \mathcal{T}_\tau^{\pi_1} v_\tau^{\pi_2}) \\
&= \langle \pi_\tau^*(\cdot|s) - \pi_t(\cdot|s), q_\tau^{\pi_t}(s,\cdot)\rangle \\
&\quad - \tau\left[\langle \pi_\tau^*(\cdot|s), \ln(\pi_\tau^*(\cdot|s))\rangle - \langle \pi_t(\cdot|s), \ln(\pi_t(\cdot|s))\rangle\right] \\
&= \langle \pi_\tau^*(\cdot|s) - \pi_t(\cdot|s), q_\tau^{\pi_t}(s,\cdot)\rangle + \tau\,\left[\mathcal{H}(\pi_\tau^*(\cdot|s)) - \mathcal{H}(\pi_t(\cdot|s))\right] \\
&\qquad\qquad \text{(By definition of } \mathcal{H}(\pi(\cdot|s))) \\
&= \langle \pi_\tau^*(\cdot|s) - \pi_t(\cdot|s), q_\zeta^t(s,\cdot)\rangle + \tau\left[\mathcal{H}(\pi_\tau^*(\cdot|s)) - \mathcal{H}(\pi_t(\cdot|s))\right] \\
&\quad + \langle \pi_\tau^*(\cdot|s) - \pi_t(\cdot|s), q_\tau^{\pi_t}(s,\cdot) - q_\zeta^t(s,\cdot)\rangle \\
&\qquad\qquad \text{(Add/Subtract } \langle \pi_\tau^*(\cdot|s) - \pi_t(\cdot|s), q_\zeta^t(s,\cdot)\rangle) \\
&\le \langle \pi_\tau^*(\cdot|s) - \pi_t(\cdot|s), q_\zeta^t(s,\cdot)\rangle + \tau\,\left[\mathcal{H}(\pi_\tau^*(\cdot|s)) - \mathcal{H}(\pi_t(\cdot|s))\right] \\
&\quad + \|\pi_\tau^*(\cdot|s) - \pi_t(\cdot|s)\|_1 \left\|q_\tau^{\pi_t}(s,\cdot) - q_\zeta^t(s,\cdot)\right\|_\infty \\
&\qquad\qquad \text{(By Holder's inequality)} \\
&\le \langle \pi_\tau^*(\cdot|s) - \pi_t(\cdot|s), q_\zeta^t(s,\cdot)\rangle + \tau\,\left[\mathcal{H}(\pi_\tau^*(\cdot|s)) - \mathcal{H}(\pi_t(\cdot|s))\right] \\
&\quad + 2\left\|q_\tau^{\pi_t}(s,\cdot) - q_\zeta^t(s,\cdot)\right\|_\infty \qquad \text{(Since } \|\pi_\tau^*(\cdot|s) - \pi_t(\cdot|s)\|_1 \le 2)
\end{aligned}$$

Define $\text{Regret}(K,u,s) = \sum_{t=0}^{K-1}\left[\langle u(\cdot|s) - \pi_t(\cdot|s), q_\zeta^t(s,\cdot)\rangle + \tau\,\left[\mathcal{H}(u(\cdot|s)) - \mathcal{H}(\pi_t(\cdot|s))\right]\right]$ as the regret incurred for state s when the comparator is policy u. Hence,

$$\begin{aligned}
&\sum_{t=0}^{K-1}\left[\mathcal{T}_\tau^{\pi_\tau^*} v_\tau^{\pi_t} - \mathcal{T}_\tau^{\pi_t} v_\tau^{\pi_t}\right](s) \le \text{Regret}(K,\pi_\tau^*,s) + 2\sum_{t=0}^{K}\left\|q_\tau^{\pi_t}(s,\cdot) - q_\zeta^t(s,\cdot)\right\|_\infty \\
&\left\|\sum_{t=0}^{K-1}\left[\mathcal{T}_\tau^{\pi_\tau^*} v_\tau^{\pi_t} - \mathcal{T}_\tau^{\pi_t} v_\tau^{\pi_t}\right]\right\|_\infty \le \max_s \text{Regret}(K,\pi_\tau^*,s) + 2\sum_{t=0}^{K}\|\boldsymbol{\epsilon}_t\|_\infty \qquad \text{(By definition of } \boldsymbol{\epsilon}_t)
\end{aligned}$$

Using the definition of $\text{Regret}(K) = \left[\text{Regret}(K,\pi_\tau^*,s_i)\right]_{i=1}^S \in \mathbb{R}^S$,

$$\left\|v_\tau^* - v_\tau^{\bar{\pi}_K}\right\|_\infty \le \frac{\|\text{Regret}(K)\|_\infty}{K\,(1-\gamma)} + \frac{2\,\sum_{t\in[K]}\|\boldsymbol{\epsilon}_t\|_\infty}{K\,(1-\gamma)}$$

□

In the special case $(0,\zeta)$, i.e., no entropy regularization from the actor side, the problem reduces to online linear optimization, as in Politex [Abbasi-Yadkori et al., 2019], yielding the following corollary.

Corollary 3 (Generic Reduction without Actor Entropy)**.** *If π^* is the optimal policy whose value function is equal to v^*, for an estimate of $q_\tau^{\pi_t}$ at iteration t s.t. $\boldsymbol{\epsilon}_t = q_\zeta^t - q_\tau^{\pi_t}$ and for any sequence of policies $\{\pi_0,\pi_1,\ldots,\pi_{K-1}\}$, if $\bar{\pi}_K$ is the corresponding mixture policy, then,*

$$\left\|v^* - v^{\bar{\pi}_K}\right\|_\infty \le \frac{\|\mathit{Regret}(K)\|_\infty}{K\,(1-\gamma)} + \frac{2\,\sum_{t\in[K]}\|\boldsymbol{\epsilon}_t\|_\infty}{K\,(1-\gamma)}\,,$$

where $(Regret(K))(s) := \sum_{t=0}^{K-1} \left[\langle \pi^*(\cdot|s) - \pi_t(\cdot|s), q_\zeta^t(s,\cdot)\rangle\right]$ *is the regret incurred on an online linear optimization problem for each state* $s \in \mathcal{S}$.

B.3.2 Proof of Corollary 1

Generic Policy Evaluation

Theorem 11 (Generic policy evaluation)**.** *Using the policy evaluation scheme in Equation* (3.1)*, if* $\delta(\tau,\zeta) := \frac{|\tau-\zeta|\ln(A)}{1-\gamma}$*, and* $TV_i := \|\pi_i(\cdot|s) - \pi_{i-1}(\cdot|s)\|_1$*, then, for all* $t \in [K]$*,*

$$\begin{aligned}\epsilon_t := \|\boldsymbol{\epsilon}_t\|_\infty &= \left\|q_\tau^{\pi_t} - q_\zeta^t\right\|_\infty \\ &\leq \frac{H_\tau\,\gamma^m}{1-\gamma}\sum_{i=1}^{t}(\gamma^m)^{t-i}\max_s\left[TV_i + \tau(1-\gamma)\left|\mathcal{H}(\pi_i(\cdot|s)) - \mathcal{H}(\pi_{i-1}(\cdot|s))\right|\right] \\ &\quad + \sum_{i=0}^{t}(\gamma^m)^{t-i}\,\delta(\tau,\zeta)\end{aligned}$$

Furthermore, if for all $i \in [t]$*,* $TV_i \leq \frac{1}{2}$*, then, for any constant* $C \in (0, 1/2)$*,*

$$\begin{aligned}\epsilon_t &\leq \frac{H_\tau\,\gamma^m}{(1-\gamma)}\sum_{i=1}^{t}(\gamma^m)^{t-i}\max_s\left[TV_i + \tau(1-\gamma)\left[TV_i\ln\left(\frac{A}{C}\right) + \left(\frac{\ln(A)}{2} + \sqrt{2}\right)\sqrt{C}\right]\right] \\ &\quad + \sum_{i=0}^{t}(\gamma^m)^{t-i}\,\delta(\tau,\zeta)\end{aligned}$$

Proof.

$$
\begin{aligned}
\epsilon_t &= \left\| q_\tau^{\pi_t} - q_\zeta^t \right\|_\infty = \left\| q_\tau^{\pi_t} - \mathcal{P}_{[0,H_\tau]}[(T_\zeta^{\pi_t})^m q_\zeta^{t-1}] \right\|_\infty && \text{(Using the update)} \\
&= \left\| \mathcal{P}_{[0,H_\tau]}[q_\tau^{\pi_t}] - \mathcal{P}_{[0,H_\tau]}[(T_\zeta^{\pi_t})^m q_\zeta^{t-1}] \right\|_\infty && \text{(Since } q_\tau^{\pi_t} \in [0, H_\tau]\text{)} \\
&= \left\| \mathcal{P}_{[0,H_\tau]}[(T_\tau^{\pi_t})^m q_\tau^{\pi_t}] - \mathcal{P}_{[0,H_\tau]}[(T_\zeta^{\pi_t})^m q_\zeta^{t-1}] \right\|_\infty && \text{(Since } q_\tau^{\pi_t} = T_\tau^{\pi_t} q_\tau^{\pi_t}\text{)} \\
&\leq \left\| (T_\tau^{\pi_t})^m q_\tau^{\pi_t} - (T_\zeta^{\pi_t})^m q_\zeta^{t-1} \right\|_\infty && \text{(Since projections are non-expansive)} \\
&\leq \left\| (T_\tau^{\pi_t})^m q_\tau^{\pi_t} - (T_\tau^{\pi_t})^m q_\zeta^{t-1} \right\|_\infty + \left\| (T_\tau^{\pi_t})^m q_\zeta^{t-1} - (T_\zeta^{\pi_t})^m q_\zeta^{t-1} \right\|_\infty \\
& && \text{(Add/Subtract } (T_\tau^{\pi_t})^m q_\zeta^{t-1} \text{ and using triangle inequality)} \\
&\leq \left\| (T_\tau^{\pi_t})^m q_\tau^{\pi_t} - (T_\tau^{\pi_t})^m q_\zeta^{t-1} \right\|_\infty + \underbrace{\frac{|\zeta - \tau|}{1-\gamma} \ln(A)}_{:=\delta(\tau,\zeta)} && \text{(Using Lemma 10)} \\
&= \left\| (T_\tau^{\pi_t})^m q_\tau^{\pi_t} - (T_\tau^{\pi_t})^m q_\zeta^{t-1} \right\|_\infty + \delta(\tau, \zeta) \\
&\leq \gamma^m \left\| q_\tau^{\pi_t} - q_\zeta^{t-1} \right\|_\infty + \delta(\tau, \zeta) && \text{(Since } T_\tau^\pi \text{ is a } \gamma \text{ contraction)} \\
&\leq \gamma^m \left\| q_\tau^{\pi_t} - q_\tau^{\pi_{t-1}} \right\|_\infty + \gamma^m \left\| q_\tau^{\pi_{t-1}} - q_\zeta^{t-1} \right\|_\infty + \delta(\tau, \zeta) \\
& && \text{(Add/subtract } q_\tau^{\pi_{t-1}} \text{ and using triangle inequality)} \\
&= \gamma^m \left\| q_\tau^{\pi_t} - q_\tau^{\pi_{t-1}} \right\|_\infty + \gamma^m \epsilon_{t-1} + \delta(\tau, \zeta)
\end{aligned}
$$

The first term is the difference in the q_τ functions between two consecutive policies, and we bound it next. For all (s, a),

$$
\begin{aligned}
q_\tau^{\pi_t}(s,a) - q_\tau^{\pi_{t-1}}(s,a) &= \gamma \sum_{s'} \mathcal{P}(s'|s,a) v_\tau^{\pi_t}(s') - \gamma \sum_{s'} \mathcal{P}(s'|s,a) v_\tau^{\pi_{t-1}}(s') \\
&= \gamma \, \mathbb{E}_{s' \sim \mathcal{P}(\cdot|s,a)}[v_\tau^{\pi_t}(s') - v_\tau^{\pi_{t-1}}(s')] \\
\implies \left\| q_\tau^{\pi_t} - q_\tau^{\pi_{t-1}} \right\|_\infty &\leq \gamma \left\| v_\tau^{\pi_t} - v_\tau^{\pi_{t-1}} \right\|_\infty
\end{aligned}
$$

Let us now bound the difference in the v_τ functions between two consecutive policies.

$$
\begin{aligned}
\left\| v_\tau^{\pi_t} - v_\tau^{\pi_{t-1}} \right\|_\infty &= \left\| T_\tau^{\pi_t} v_\tau^{\pi_t} - T_\tau^{\pi_t} v_\tau^{\pi_{t-1}} + T_\tau^{\pi_t} v_\tau^{\pi_{t-1}} - T_\tau^{\pi_{t-1}} v_\tau^{\pi_{t-1}} \right\|_\infty \\
& && \text{(Since } T_\tau^\pi v_\tau^\pi = v_\tau^\pi \text{ and add/subtract } T_\tau^{\pi_t} v_\tau^{\pi_{t-1}}\text{)} \\
&\leq \left\| T_\tau^{\pi_t} v_\tau^{\pi_t} - T_\tau^{\pi_t} v_\tau^{\pi_{t-1}} \right\|_\infty + \left\| T_\tau^{\pi_t} v_\tau^{\pi_{t-1}} - T_\tau^{\pi_{t-1}} v_\tau^{\pi_{t-1}} \right\|_\infty \\
& && \text{(Triangle inequality)} \\
&\leq \gamma \left\| v_\tau^{\pi_t} - v_\tau^{\pi_{t-1}} \right\|_\infty + \left\| T_\tau^{\pi_t} v_\tau^{\pi_{t-1}} - T_\tau^{\pi_{t-1}} v_\tau^{\pi_{t-1}} \right\|_\infty \\
& && \text{(Since } T_\tau^\pi \text{ is a } \gamma\text{-contraction)} \\
\implies \left\| v_\tau^{\pi_t} - v_\tau^{\pi_{t-1}} \right\|_\infty &\leq \frac{1}{1-\gamma} \left\| T_\tau^{\pi_t} v_\tau^{\pi_{t-1}} - T_\tau^{\pi_{t-1}} v_\tau^{\pi_{t-1}} \right\|_\infty
\end{aligned}
$$

In order to bound $\left\|T_\tau^{\pi_t} v_\tau^{\pi_{t-1}} - T_\tau^{\pi_{t-1}} v_\tau^{\pi_{t-1}}\right\|_\infty$, consider a fixed state s. By definition of T_τ^π,

$$\begin{aligned}
T_\tau^{\pi_t} v_\tau^{\pi_{t-1}}(s) - T_\tau^{\pi_{t-1}} v_\tau^{\pi_{t-1}}(s) &= \langle \pi_t(\cdot|s) - \pi_{t-1}(\cdot|s), r(s,\cdot)\rangle \\
&\quad + \gamma \sum_a [\pi_t(a|s) - \pi_{t-1}(a|s)]\, \mathbb{E}_{s'\sim\mathcal{P}(\cdot|s,a)} v_\tau^{\pi_{t-1}}(s') \\
&\quad + \tau[\mathcal{H}(\pi_t(\cdot|s)) - \mathcal{H}(\pi_{t-1}(\cdot|s))] \\
&\leq \|\pi_t(\cdot|s) - \pi_{t-1}\|_1 \, \|r(s,\cdot)\|_\infty \\
&\quad + \gamma \, \|\pi_t(\cdot|s) - \pi_{t-1}\|_1 \, \|v_\tau^{\pi_{t-1}}\|_\infty \\
&\quad + \tau[\mathcal{H}(\pi_t(\cdot|s)) - \mathcal{H}(\pi_{t-1}(\cdot|s))] \text{ (By Holder's inequality)} \\
&\leq (1 + \gamma\, H_\tau) \, \|\pi_t(\cdot|s) - \pi_{t-1}(\cdot|s)\|_1 \\
&\quad + \tau|\mathcal{H}(\pi_t(\cdot|s)) - \mathcal{H}(\pi_{t-1}(\cdot|s))| \\
&\qquad \text{(Since rewards are in } [0,1] \text{ and } v_\tau^{\pi_{t-1}}(s) \leq H_\tau) \\
&\leq H_\tau \, \|\pi_t(\cdot|s) - \pi_{t-1}(\cdot|s)\|_1 + \tau|\mathcal{H}(\pi_t(\cdot|s)) - \mathcal{H}(\pi_{t-1}(\cdot|s))| \\
&\qquad \text{(Since } 1 \leq 1 + \tau \ln(A) = (1-\gamma)\, H_\tau) \\
\implies \left\|T_\tau^{\pi_t} v_\tau^{\pi_{t-1}} - T_\tau^{\pi_{t-1}} v_\tau^{\pi_{t-1}}\right\|_\infty &\leq H_\tau \, \Big(\max_s \left[TV_i + \tau(1-\gamma)\, |\mathcal{H}(\pi_t(\cdot|s)) - \mathcal{H}(\pi_{t-1}(\cdot|s))|\right]\Big) \\
&\qquad \text{(Since } 1 \leq (1-\gamma)\, H_\tau)
\end{aligned}$$

Combining the above inequalities,

$$\begin{aligned}
\|v_\tau^{\pi_t} - v_\tau^{\pi_{t-1}}\|_\infty &\leq \frac{H_\tau}{(1-\gamma)} \, \max_s \left[TV_i + \tau(1-\gamma)\, |\mathcal{H}(\pi_t(\cdot|s)) - \mathcal{H}(\pi_{t-1}(\cdot|s))|\right] \\
\implies \|q_\tau^{\pi_t} - q_\tau^{\pi_{t-1}}\|_\infty &\leq \frac{H_\tau}{(1-\gamma)} \, \max_s \left[TV_i + \tau(1-\gamma)\, |\mathcal{H}(\pi_t(\cdot|s)) - \mathcal{H}(\pi_{t-1}(\cdot|s))|\right]
\end{aligned}$$

$$\begin{aligned}
\implies \epsilon_t &\leq \underbrace{\frac{H_\tau\, \gamma^m}{(1-\gamma)} \, \max_s \left[TV_i + \tau(1-\gamma)\, |\mathcal{H}(\pi_t(\cdot|s)) - \mathcal{H}(\pi_{t-1}(\cdot|s))|\right]}_{:=B_t} + \gamma^m\, \epsilon_{t-1} + \delta(\tau,\zeta) \\
\implies \epsilon_t &\leq B_t + \gamma^m\, \epsilon_{t-1} + \delta(\tau,\zeta)
\end{aligned}$$

Bounding ϵ_0,

$$\epsilon_0 = q_\tau^{\pi_0} - q_\zeta^{\pi_0} = \mathbb{E}_{s'\sim\mathcal{P}(\cdot|s,a)}\left[r(s,a) + \gamma v_\tau^{\pi}(s')\right] - \mathbb{E}_{s'\sim\mathcal{P}(\cdot|s,a)}\left[r(s,a) + \gamma v_\zeta^{\pi}(s')\right] \quad \text{(By definition)}$$

$$= \gamma\mathbb{E}_{s'\sim\mathcal{P}(\cdot|s,a)}[v_\tau^{\pi_0}(s') - v_\zeta^{\pi_0}(s')]$$

$$= \gamma\mathbb{E}_{s'\sim\mathcal{P}(\cdot|s,a)}\Bigg[[v^{\pi_0}(s') + \tau\sum_{t=0}^{\infty}\gamma^t\left[\mathcal{H}(\pi_0(\cdot|s_t))|s_0 = s'\right] - [v^{\pi_0}(s') + \zeta\sum_{t=0}^{\infty}\gamma^t\left[\mathcal{H}(\pi_0(\cdot|s_t))|s_0 = s'\right]]\Bigg] \quad \text{(By definition)}$$

$$\leq \frac{|\tau - \zeta|\,\ln(A)}{1-\gamma} = \delta(\tau,\zeta) \quad \text{(Since } \mathcal{H}(\pi(\cdot|s)) \leq \ln(A)\text{)}$$

For a fixed $t \in [K]$, recursing from $i = t-1$ to $i = 1$ and using that

$$\epsilon_t \leq (\gamma^m)^t\epsilon_0 + \sum_{i=1}^{t}(\gamma^m)^{t-i}\left(B_i + \delta(\tau,\zeta)\right)$$

$$\leq \frac{H_\tau\,\gamma^m}{(1-\gamma)}\sum_{i=1}^{t}(\gamma^m)^{t-i}\max_s\left[TV_i + \tau(1-\gamma)\left|\mathcal{H}(\pi_i(\cdot|s)) - \mathcal{H}(\pi_{i-1}(\cdot|s))\right|\right] + \sum_{i=0}^{t}(\gamma^m)^{t-i}\,\delta(\tau,\zeta)$$

Furthermore, if $TV_i \leq \frac{1}{2}$ for all $i \in [t]$, using Lemma 7, we can further upper-bound $|\mathcal{H}(\pi_i(\cdot|s)) - \mathcal{H}(\pi_{i-1}(\cdot|s))|$ to get that for any constant $C \in (0, 1/2)$,

$$|\mathcal{H}(\pi_i(\cdot|s)) - \mathcal{H}(\pi_{i-1}(\cdot|s))| \leq TV_i\,\ln\left(\frac{A}{C}\right) + \left(\frac{\ln(A-1)}{2} + \sqrt{2}\right)\sqrt{C}$$

Combining the above relations, in this case, we get that,

$$\epsilon_t \leq \frac{H_\tau\,\gamma^m}{(1-\gamma)}\sum_{i=1}^{t}(\gamma^m)^{t-i}\max_s\left[TV_i + \tau(1-\gamma)\left[TV_i\,\ln\left(\frac{A}{C}\right) + \left(\frac{\ln(A)}{2} + \sqrt{2}\right)\sqrt{C}\right]\right] + \sum_{i=0}^{t}(\gamma^m)^{t-i}\,\delta(\tau,\zeta)$$

□

Corollary 4. *Using the policy evaluation scheme in Equation* (3.1) *with* $\zeta = \tau$, *for all* $t \in [K]$, *if for all* $i \in [t]$, $TV_i := \|\pi_i(\cdot|s) - \pi_{i-1}(\cdot|s)\|_1 \leq \frac{1}{2}$, *then, for any constant* $C \in (0, 1/2)$,

$$\epsilon_t \leq \frac{H_\tau\,\gamma^m}{(1-\gamma)}\sum_{i=1}^{t}(\gamma^m)^{t-i}\max_s\left[TV_i + \tau(1-\gamma)\left[TV_i\,\ln\left(\frac{A}{C}\right) + \left(\frac{\ln(A)}{2} + \sqrt{2}\right)\sqrt{C}\right]\right]$$

Proof. Setting $\zeta = \tau$ in Theorem 11. □

Corollary 5. *Using the policy evaluation scheme in Equation* (3.1) *with* $\zeta = 0$*, for all* $t \in [K]$*, if for all* $i \in [t]$*,* $TV_i := \|\pi_i(\cdot|s) - \pi_{i-1}(\cdot|s)\|_1 \leq \frac{1}{2}$*, then, for any constant* $C \in (0, 1/2)$*,*

$$\epsilon_t \leq \frac{H_\tau\, \gamma^m}{(1-\gamma)} \sum_{i=1}^{t} (\gamma^m)^{t-i} \max_s \left[TV_i + \tau(1-\gamma) \left[TV_i \ln\left(\frac{A}{C}\right) + \left(\frac{\ln(A)}{2} + \sqrt{2}\right) \sqrt{C} \right] \right] + \frac{\tau \ln(A)}{(1-\gamma)^2}$$

Proof. Setting $\zeta = 0$ in Theorem 11. □

The following proposition shows that the objective in Equation (3.5b) admits a closed-form solution. Substituting the intermediate policies NPG and SPMA into the closed-form expression in Equation (B.1) yields their entropy-regularized (actor entropy) counterparts, referred to as soft NPG and soft SPMA.

Proposition 1. *If* $\alpha_t := \frac{1}{1+\tau_t}$*, the closed-form solution for the proximal update in Equation* (3.5b) *for any* s, a *is given as,*

$$\pi_{t+1}(a|s) = \frac{[\pi_{t+1/2}(a|s)]^{\alpha_t}}{\sum_{a'} [\pi_{t+1/2}(a'|s)]^{\alpha_t}} \tag{B.1}$$

Proof.

$$\begin{aligned}
\pi_{t+1}(\cdot|s) &= \arg\min_{\pi(\cdot|s)\in\Delta} \Big[\mathrm{KL}(\pi(\cdot|s)||\pi_{t+1/2}(\cdot|s)) - \tau_t \mathcal{H}(\pi(\cdot|s)) \Big] \\
&= \arg\min_{\pi(\cdot|s)\in\Delta} \mathbb{E}_{a\sim\pi(.|s)} \Big[(1+\tau_t) \ln(\pi(.|s)) - \ln(\pi_{t+1/2}(\cdot|s)) \Big] \\
&\qquad\qquad \text{(Using the definition of } H(\pi(\cdot|s))) \\
&= \arg\min_{\pi(\cdot|s)\in\Delta} \mathbb{E}_{a\sim\pi(.|s)} \Big[\ln(\pi(.|s)) - \alpha_t \ln(\pi_{t+1/2}(\cdot|s)) \Big] \qquad \text{(Since } \alpha_t = \tfrac{1}{1+\tau_t}) \\
&= \arg\min_{\pi(\cdot|s)\in\Delta} \mathbb{E}_{a\sim\pi(.|s)} \Big[\ln(\pi(.|s)) - \ln([\pi_{t+1/2}(\cdot|s)]^{\alpha_t}) \Big] \\
&= \arg\min_{\pi(\cdot|s)\in\Delta} \mathrm{KL}(\pi(.|s)||\pi_{t+1/2}(\cdot|s)^{\alpha_t}) \qquad \text{(By definition of the KL divergence)}
\end{aligned}$$

Using the fact that KL projection onto the simplex results in normalization, we get that,

$$\pi_{t+1}(a|s) = \frac{[\pi_{t+1/2}(a|s)]^{\alpha_t}}{\sum_{a'} [\pi_{t+1/2}(a'|s)]^{\alpha_t}}$$

□

Note that when $\tau = 0$, we have $\tau_t = 0$ and $\alpha_t = 1$ for all t, recovering the standard unregularized updates for both NPG and SPMA. For $\tau > 0$, the soft updates can be expressed as: for any s, a, with $\alpha_t = \frac{1}{1+\tau_t}$ and $\tau_t = \eta_t \tau$,

$$\pi_{t+1}(a|s) = \frac{[\pi_t(a|s)]^{\alpha_t} \exp(\eta_t \alpha_t q_\zeta^t(s,a))}{\mathcal{Z}_t}, \tag{B.2}$$

$$\text{with} \quad \mathcal{Z}_t = \sum_{a'} [\pi_t(a'|s)]^{\alpha_t} \exp(\eta_t \alpha_t q_\zeta^t(s,a')), \qquad \textbf{(Soft-NPG)}$$

$$\pi_{t+1}(a|s) = \frac{[\pi_t(a|s)]^{\alpha_t} \left[1 + \eta_t \left(q_\zeta^t(s,a) - v_\zeta^t(s)\right)\right]^{\alpha_t}}{\mathcal{Z}_t}, \tag{B.3}$$

$$\text{with} \quad \mathcal{Z}_t = \sum_{a'} [\pi_t(a'|s)]^{\alpha_t} \left[1 + \eta_t \left(q_\zeta^t(s,a') - v_\zeta^t(s)\right)\right]^{\alpha_t}, \qquad \textbf{(Soft-SPMA)}$$

Policy Error Bound for (soft) NPG

In the next corollary, we use Theorem 11 with $\zeta = \tau$ to instantiate the policy error bound for soft NPG with entropy-regularized policy evaluation.

Corollary 6 (Policy evaluation with $\zeta = \tau$)**.** *Using the policy evaluation update in Equation* (3.1) *with* $\zeta = \tau$*, for soft* NPG *with* $\eta_t = \frac{1}{c+\tau\,(t+1)}$ *and a constant* $c \geq \max\left\{\frac{8\,(1+\tau\,\ln(A))}{(1-\gamma)}, 32\,\tau\,\ln(A)\right\}$*, for all* $t \in [K]$*,* $\epsilon_t := \|\boldsymbol{\epsilon}_t\|_\infty$ *can be bounded as:*

$$\epsilon_t \leq \frac{8(1+\tau\ln(A))\gamma^m}{(1-\gamma)^3} \Bigg[\left(1 + \tau \ln\left(A\,K^4\right)\right) \left(\left(\ln(A\,K^4) + \frac{1+\tau\ln(A)}{\tau(1-\gamma)}\right) \left(\frac{1}{t} + (\gamma^m)^{t/2}\right) + \frac{\sqrt{\tau}}{K} \right) + \tau\ (\ln(A)+1)\ \frac{1}{K^2} \Bigg]$$

Proof. For a fixed iteration $t \in [K]$ and state $s \in \mathcal{S}$, let us first bound $\|\pi_{t+1}(\cdot|s) - \pi_t(\cdot|s)\|_1$.

$$\|\pi_{t+1}(\cdot|s) - \pi_t(\cdot|s)\|_1 \leq \left\|\pi_{t+1}(\cdot|s) - \pi_{t+1/2}(\cdot|s)\right\|_1 + \left\|\pi_{t+1/2}(\cdot|s) - \pi_t(\cdot|s)\right\|_1$$
(Triangle inequality)

We first bound $\left\|\pi_{t+1/2}(\cdot|s) - \pi_t(\cdot|s)\right\|_1$. Using the mirror descent view of NPG [Xiao, 2022], the update can be written as:

$$\begin{aligned}
&\pi_{t+1/2}(\cdot|s) = \arg\min_{\pi\in\Delta} \Big[-\eta_t \langle q_\tau^t(s,\cdot), \pi(\cdot|s)\rangle + \mathrm{KL}(\pi(\cdot|s)||\pi_t(\cdot|s))\Big] \\
\implies &-\eta_t \langle q_\tau^t(s,\cdot), \pi_{t+1/2}(\cdot|s)\rangle + \mathrm{KL}(\pi_{t+1/2}(\cdot|s)||\pi_t(\cdot|s)) \le -\eta_t \langle q_\tau^t(s,\cdot), \pi_t(\cdot|s)\rangle \\
&\qquad\qquad + \mathrm{KL}(\pi_t(\cdot|s)||\pi_t(\cdot|s)) \\
\implies &\frac{1}{2}\left\|\pi_{t+1/2}(\cdot|s) - \pi_t(\cdot|s)\right\|_1^2 \le \mathrm{KL}(\pi_{t+1/2}(\cdot|s)||\pi_t(\cdot|s)) \le \eta_t \langle q_\tau^t(s,\cdot), \pi_{t+1/2}(\cdot|s) - \pi_t(\cdot|s)\rangle \\
&\qquad\qquad \text{(By Pinsker's inequality)} \\
&\qquad\qquad \le \eta_t \left\|q_\tau^t(s,\cdot)\right\|_\infty \left\|\pi_{t+1/2}(\cdot|s) - \pi_t(\cdot|s)\right\|_1 \\
&\qquad\qquad \text{(By Holder's inequality)} \\
\implies &\left\|\pi_{t+1/2}(\cdot|s) - \pi_t(\cdot|s)\right\|_1 \le 2\eta_t H_\tau \qquad \text{(Since } \left\|q_\tau^t(s,\cdot)\right\|_\infty \le H_\tau\text{)}
\end{aligned}$$

In order to bound $\left\|\pi_{t+1}(\cdot|s) - \pi_{t+1/2}(\cdot|s)\right\|_1$, we use the Equation (3.5b) update. Specifically,

$$\pi_{t+1}(\cdot|s) = \arg\min_{\pi\in\Delta} \Big[\mathrm{KL}(\pi(\cdot|s)||\pi_{t+1/2}(\cdot|s)) - \tau_t \mathcal{H}(\pi(\cdot|s))\Big]$$

$$\begin{aligned}
\implies \mathrm{KL}(\pi_{t+1}(\cdot|s)||\pi_{t+1/2}(\cdot|s)) - \tau_t \mathcal{H}(\pi_{t+1}(\cdot|s)) &\le \mathrm{KL}(\pi_{t+1/2}(\cdot|s)||\pi_{t+1/2}(\cdot|s)) \\
&\quad - \tau_t \mathcal{H}(\pi_{t+1/2}(\cdot|s)) \\
\frac{1}{2}\left\|\pi_{t+1}(\cdot|s) - \pi_{t+1/2}(\cdot|s)\right\|_1^2 &\le \mathrm{KL}(\pi_{t+1}||\pi_{t+1/2}(\cdot|s)) \\
&\qquad \text{(By Pinsker's inequality)} \\
&\le \tau_t \mathcal{H}(\pi_{t+1}(\cdot|s)) - \tau_t \mathcal{H}(\pi_{t+1/2}(\cdot|s)) \\
&\le \tau_t \ln(A) \\
\implies \left\|\pi_{t+1}(\cdot|s) - \pi_{t+1/2}(\cdot|s)\right\|_1 &\le \sqrt{2\,\tau\,\eta_t \ln(A)} \qquad \text{(since } \tau_t = \tau\,\eta_t\text{)}
\end{aligned}$$

Combining the above relations,

$$\|\pi_{t+1}(\cdot|s) - \pi_t(\cdot|s)\|_1 \le \sqrt{2\,\tau\,\eta_t \ln(A)} + 2\,\eta_t H_\tau \tag{B.4}$$

Using Equation (B.4) in Theorem 11, for the special case $\|\pi_{t+1}(\cdot|s) - \pi_t(\cdot|s)\|_1 \le \frac{1}{2}$ for all $t \in [K]$, η_t must satisfy the following conditions:

- $2\,\eta_t H_\tau \le \frac{1}{4} \implies \eta_t \le \frac{(1-\gamma)}{8\,(1+\tau\,\ln(A))}$
- $\sqrt{2\,\tau\,\eta_t\,\ln(A)} \le \frac{1}{4} \implies \eta_t \le \frac{1}{32\,\tau\,\ln(A)}$

For $\eta_t = \frac{1}{c+\tau(t+1)} \le \frac{1}{c}$, it is thus sufficient to ensure that,

$$c \ge \max\left\{\frac{8\,(1+\tau\,\ln(A))}{(1-\gamma)}, 32\,\tau\,\ln(A)\right\}$$

Given this constraint on c, we use Theorem 11 with $\zeta = \tau$ and $TV_i := \|\pi_i(\cdot|s) - \pi_{i-1}(\cdot|s)\|_1$, to get that, for any constant $C \in (0, 1/2)$, for all $t \in [K]$,

$$\epsilon_t \leq \frac{H_\tau\, \gamma^m}{(1-\gamma)} \sum_{i=1}^{t} (\gamma^m)^{t-i} \max_s \left[TV_i + \tau(1-\gamma) \left[TV_i \ln\left(\frac{A}{C}\right) + \left(\frac{\ln(A)}{2} + \sqrt{2}\right) \sqrt{C} \right] \right]$$

Observe that the choice of c ensures $\|\pi_{t+1}(\cdot|s) - \pi_t(\cdot|s)\|_1 \leq \frac{1}{2}$. Using Lemma 8 we obtain,

$$\|\pi_{t+1}(\cdot|s) - \pi_t(\cdot|s)\|_1 \leq 4\,\tau\,\eta_t \ln(\frac{A}{C}) + 2\sqrt{\tau}\, C^{\frac{1}{4}} + 2\,\eta_t\, H_\tau$$

Combining the above with the upper bound on ϵ_t, we get,

$$\epsilon_t \leq \frac{H_\tau\, \gamma^m}{(1-\gamma)} \sum_{i=1}^{t} (\gamma^m)^{t-i} \Bigg[\left(1 + \tau\,(1-\gamma)\, \ln(\frac{A}{C})\right) \left(2\,\eta_i\,(2\,\tau\, \ln(\frac{A}{C}) + H_\tau) + 2\sqrt{\tau}\, C^{\frac{1}{4}}\right) + \tau(1-\gamma) \left(\frac{\ln(A)}{2} + \sqrt{2}\right) \sqrt{C} \Bigg]$$

Setting $C = \frac{1}{K^4}$,

$$\leq \frac{H_\tau\, \gamma^m}{(1-\gamma)} \Bigg[2 \left(1 + \tau \ln\left(A\, K^4\right)\right) \left(\left(2\ln(A\, K^4) + \frac{H_\tau}{\tau}\right) \sum_{i=1}^{t} \frac{(\gamma^m)^{t-i}}{(i+1)} + \frac{\sqrt{\tau}}{K(1-\gamma^m)} \right) + \frac{\tau}{1-\gamma^m} \left(\frac{\ln(A)}{2} + \sqrt{2}\right) \frac{1}{K^2} \Bigg] \qquad \text{(since } \gamma < 1\text{)}$$

Using Lemma 9, we can bound $\sum_{i=1}^{t} \frac{(\gamma^m)^{t-i}}{(i+1)} \leq \frac{2}{1-\gamma^m} \left(\frac{1}{t} + (\gamma^m)^{t/2}\right) \leq \frac{2}{1-\gamma} \left(\frac{1}{t} + (\gamma^m)^{t/2}\right)$.

$$\implies \epsilon_t \leq \frac{H_\tau\, \gamma^m}{(1-\gamma)} \Bigg[4 \left(1 + \tau \ln\left(A\, K^4\right)\right) \left(\left(2\ln(A\, K^4) + \frac{H_\tau}{\tau}\right) \left(\frac{1}{t} + (\gamma^m)^{t/2}\right) + \frac{\sqrt{\tau}}{K(1-\gamma^m)} \right) + \frac{\tau}{1-\gamma^m} \left(\frac{\ln(A)}{2} + \sqrt{2}\right) \frac{1}{K^2} \Bigg]$$

$$\leq \frac{8 H_\tau\, \gamma^m}{(1-\gamma)^2} \Bigg[\left(1 + \tau \ln\left(A\, K^4\right)\right) \left(\left(\ln(A\, K^4) + \frac{H_\tau}{\tau}\right) \left(\frac{1}{t} + (\gamma^m)^{t/2}\right) + \frac{\sqrt{\tau}}{K} \right) + \tau\, (\ln(A) + 1)\, \frac{1}{K^2} \Bigg] \qquad \text{(Since } \gamma < 1\text{)}$$

□

In the next corollary, we use Theorem 11 with $\zeta = 0$ to instantiate the policy error bound for (soft) NPG with entropy-regularized policy evaluation.

Corollary 7 (Policy evaluation with $\zeta = 0$)**.** *Using the policy evaluation update in Equation* (3.1) *with* $\zeta = 0$, $\epsilon_t := \|\boldsymbol{\epsilon}_t\|_\infty$ *can be bounded as:*

- ***Soft*** **NPG***: If* $\eta_t = \frac{1}{c+\tau\,(t+1)}$ *for a constant* $c \geq \max\left\{\frac{8\,(1+\tau\,\ln(A))}{(1-\gamma)}, 32\,\tau\,\ln(A)\right\}$*, then, for all* $t \in [K]$*,*

$$\epsilon_t \leq \frac{8(1+\tau\ln(A))\gamma^m}{(1-\gamma)^3}\left[\left(1+\tau\ln\left(A\,K^4\right)\right)\right.$$
$$\left(\left(\ln(A\,K^4)+\frac{1+\tau\ln(A)}{\tau(1-\gamma)}\right)\left(\frac{1}{t}+(\gamma^m)^{t/2}\right)+\frac{\sqrt{\tau}}{K}\right)$$
$$\left.+\,\tau\,(\ln(A)+1)\,\frac{1}{K^2}\right]+\frac{\tau}{(1-\gamma)^2}\ln(A)$$

- **NPG***: If* $\eta_t = \eta = \frac{(1-\gamma)\sqrt{\ln(A)}}{\sqrt{K}}$*, then, for all* $t \in [K]$*,*

$$\epsilon_t \leq \frac{2\sqrt{\ln(A)}\,\gamma^m}{(1-\gamma)^3}\,\frac{1}{\sqrt{K}}$$

Proof. Using Theorem 11 for soft NPG and Corollary 5 for NPG, and following the same proof as Corollary 6. □

Policy Error Bound for (soft) SPMA

In the next corollary, we use Theorem 11 with $\zeta = \tau$ to instantiate the policy error bound for soft SPMA with entropy-regularized policy evaluation.

Corollary 8 (Policy evaluation with $\zeta = \tau$)**.** *Using the policy evaluation update in Equation* (3.1) *with* $\zeta = \tau$*, for soft* SPMA *with* $\eta_t = \frac{1}{c+\tau\,(t+1)}$ *and a constant* $c \geq \max\left\{\frac{4(1+\tau\,\ln(A))}{(1-\gamma)}, 32\,\tau\,\ln(A)\right\}$*, for all* $t \in [K]$*,* ϵ_t *can be bounded as:*

$$\epsilon_t := \|\boldsymbol{\epsilon}_t\|_\infty \leq \frac{8(1+\tau\ln(A))\gamma^m}{(1-\gamma)^3}\left[\left(1+\tau\ln\left(A\,K^4\right)\right)\right.$$
$$\left(\left(\ln(A\,K^4)+\frac{1+\tau\ln(A)}{\tau(1-\gamma)}\right)\left(\frac{1}{t}+(\gamma^m)^{t/2}\right)+\frac{\sqrt{\tau}}{K}\right)$$
$$\left.+\,\tau\,(\ln(A)+1)\,\frac{1}{K^2}\right]$$

Proof. For a fixed iteration $t \in [K]$ and state $s \in \mathcal{S}$, let us first bound $\|\pi_{t+1}(\cdot|s) - \pi_t(\cdot|s)\|_1$.

$$\begin{aligned}
\|\pi_{t+1}(\cdot|s) - \pi_t(\cdot|s)\|_1 &\leq \left\|\pi_{t+1}(\cdot|s) - \pi_{t+1/2}(\cdot|s)\right\|_1 + \left\|\pi_{t+1/2}(\cdot|s) - \pi_t(\cdot|s)\right\|_1 \\
&\qquad \text{(Triangle inequality)} \\
&\leq \left\|\pi_{t+1}(\cdot|s) - \pi_{t+1/2}(\cdot|s)\right\|_1 + \eta_t \sum_a \pi_t(a)\left[q_\tau^t(s,a) - v_\tau^t(s)\right] \\
&\qquad \text{(By the SPMA update in Equation (3.4))} \\
\implies \|\pi_{t+1}(\cdot|s) - \pi_t(\cdot|s)\|_1 &\leq \left\|\pi_{t+1}(\cdot|s) - \pi_{t+1/2}(\cdot|s)\right\|_1 + \eta_t H_\tau \\
&\qquad \text{(Since } |q_\tau^t(s,a) - v_\tau^t(s)| \leq H_\tau\text{)}
\end{aligned}$$

In order to bound $\left\|\pi_{t+1}(\cdot|s) - \pi_{t+1/2}(\cdot|s)\right\|_1$, we use the Equation (3.5b) update. Specifically,

$$\pi_{t+1}(\cdot|s) = \arg\min_{\pi \in \Delta} \left[\mathrm{KL}(\pi||\pi_{t+1/2}(\cdot|s)) - \tau_t \mathcal{H}(\pi(\cdot|s))\right]$$

$$\begin{aligned}
\implies \mathrm{KL}(\pi_{t+1}(\cdot|s)||\pi_{t+1/2}(\cdot|s)) - \tau_t \mathcal{H}(\pi_{t+1}(\cdot|s)) &\leq \mathrm{KL}(\pi_{t+1/2}(\cdot|s)||\pi_{t+1/2}(\cdot|s)) \\
&\quad - \tau_t \mathcal{H}(\pi_{t+1/2}(\cdot|s)) \\
\frac{1}{2}\left\|\pi_{t+1}(\cdot|s) - \pi_{t+1/2}(\cdot|s)\right\|_1^2 &\leq \mathrm{KL}(\pi_{t+1}(\cdot|s)||\pi_{t+1/2}(\cdot|s)) \\
&\qquad \text{(By Pinsker's inequality)} \\
&\leq \tau_t \mathcal{H}(\pi_{t+1}(\cdot|s)) - \tau_t \mathcal{H}(\pi_{t+1/2}(\cdot|s)) \\
&\leq \tau_t \ln(A) \qquad \text{(Since } \mathcal{H}(\pi) \in [0, \ln(A)]\text{)} \\
\implies \left\|\pi_{t+1}(\cdot|s) - \pi_{t+1/2}(\cdot|s)\right\|_1 &\leq \sqrt{2\,\tau\,\eta_t\,\ln(A)} \qquad \text{(Since } \tau_t = \tau\,\eta_t\text{)}
\end{aligned}$$

Combining the above relations,

$$\|\pi_{t+1}(\cdot|s) - \pi_t(\cdot|s)\|_1 \leq \sqrt{2\,\tau\,\eta_t\,\ln(A)} + \eta_t\,H_\tau \tag{B.5}$$

In order to use Theorem 11, we need to ensure that $\|\pi_{t+1}(\cdot|s) - \pi_t(\cdot|s)\|_1 \leq \frac{1}{2}$ for all $t \in [K]$. Using Equation (B.5), it is sufficient to ensure that η_t satisfies the following relations:

- $\eta_t\,H_\tau \leq \frac{1}{4} \implies \eta_t \leq \frac{(1-\gamma)}{4\,(1+\tau\,\ln(A))}$
- $\sqrt{2\,\tau\,\eta_t\,\ln(A)} \leq \frac{1}{4} \implies \eta_t \leq \frac{1}{32\,\tau\,\ln(A)}$

For $\eta_t = \frac{1}{c+\tau(t+1)} \leq \frac{1}{c}$, it is thus sufficient to ensure that,

$$c \geq \max\left\{\frac{4\,(1+\tau\,\ln(A))}{(1-\gamma)}, 32\,\tau\,\ln(A)\right\}$$

Given this constraint on c, we use Theorem 11 with $\zeta = \tau$ and $TV_i := \|\pi_i(\cdot|s) - \pi_{i-1}(\cdot|s)\|_1$, to get that, for any constant $C \in (0, 1/2)$, for all $t \in [K]$,

$$\epsilon_t \leq \frac{H_\tau\, \gamma^m}{(1-\gamma)} \sum_{i=1}^{t} (\gamma^m)^{t-i} \max_s \left[TV_i + \tau(1-\gamma) \left(TV_i \ln\left(\frac{A}{C}\right) + \left(\frac{\ln(A)}{2} + \sqrt{2}\right) \sqrt{C} \right) \right]$$

Observe that the choice of c ensures $\|\pi_{t+1}(\cdot|s) - \pi_t(\cdot|s)\|_1 \leq \frac{1}{2}$. Using Lemma 8 we obtain,

$$\|\pi_{t+1}(\cdot|s) - \pi_t(\cdot|s)\|_1 \leq 4\,\tau\,\eta_t \ln(\frac{A}{C}) + 2\sqrt{\tau}\, C^{\frac{1}{4}} + 2\,\eta_t\, H_\tau$$

Combining the above with the upper bound on ϵ_t, we get,

$$\epsilon_t \leq \frac{H_\tau\, \gamma^m}{(1-\gamma)} \sum_{i=1}^{t} (\gamma^m)^{t-i} \left[\left(1 + \tau\,(1-\gamma)\, \ln(\frac{A}{C})\right) \left(2\,\eta_i\,(2\,\tau\, \ln(\frac{A}{C}) + H_\tau) + 2\sqrt{\tau}\, C^{\frac{1}{4}}\right) \right.$$
$$\left. + \tau(1-\gamma) \left(\frac{\ln(A)}{2} + \sqrt{2}\right) \sqrt{C} \right]$$

Setting $C = \frac{1}{K^4}$,

$$\leq \frac{H_\tau\, \gamma^m}{(1-\gamma)} \left[2 \left(1 + \tau \ln\left(A\, K^4\right)\right) \left(\left(2\ln(A\, K^4) + \frac{H_\tau}{\tau}\right) \sum_{i=1}^{t} \frac{(\gamma^m)^{t-i}}{(i+1)} + \frac{\sqrt{\tau}}{K(1-\gamma^m)} \right) \right.$$
$$\left. + \frac{\tau}{1-\gamma^m} \left(\frac{\ln(A)}{2} + \sqrt{2}\right) \frac{1}{K^2} \right] \qquad \text{(since } \gamma < 1\text{)}$$

Using Lemma 9, we can bound $\sum_{i=1}^{t} \frac{(\gamma^m)^{t-i}}{(i+1)} \leq \frac{2}{1-\gamma^m} \left(\frac{1}{t} + (\gamma^m)^{t/2}\right) \leq \frac{2}{1-\gamma} \left(\frac{1}{t} + (\gamma^m)^{t/2}\right)$.

$$\implies \epsilon_t \leq \frac{H_\tau\, \gamma^m}{(1-\gamma)} \left[4 \left(1 + \tau \ln\left(A\, K^4\right)\right) \right.$$
$$\left(\left(2\ln(A\, K^4) + \frac{H_\tau}{\tau}\right) \left(\frac{1}{t} + (\gamma^m)^{t/2}\right) + \frac{\sqrt{\tau}}{K(1-\gamma^m)} \right)$$
$$\left. + \frac{\tau}{1-\gamma^m} \left(\frac{\ln(A)}{2} + \sqrt{2}\right) \frac{1}{K^2} \right]$$
$$\leq \frac{8 H_\tau\, \gamma^m}{(1-\gamma)^2} \left[\left(1 + \tau \ln\left(A\, K^4\right)\right) \left(\left(\ln(A\, K^4) + \frac{H_\tau}{\tau}\right) \left(\frac{1}{t} + (\gamma^m)^{t/2}\right) + \frac{\sqrt{\tau}}{K} \right) \right.$$
$$\left. + \tau\, (\ln(A) + 1)\, \frac{1}{K^2} \right] \qquad \text{(Since } \gamma < 1\text{)}$$

□

In the next corollary, we use Theorem 11 with $\zeta = 0$ to instantiate the policy error bound for (soft) SPMA without entropy-regularized policy evaluation.

Corollary 9 (Policy evaluation with $\zeta = 0$)**.** *Using the policy evaluation update in Equation* (3.1) *with (soft)* SPMA*,* $\epsilon_t := \|\boldsymbol{\epsilon}_t\|_\infty$ *can be bounded as:*

- ***Soft*** **SPMA***: if* $\eta_t = \frac{1}{c+\tau\,(t+1)}$ *for a constant* $c \geq \max\left\{\frac{4(1+\tau\,\ln(A))}{(1-\gamma)}, 32\,\tau\,\ln(A)\right\}$*, then, for all* $t \in [K]$*,*

$$\epsilon_t \leq \frac{8(1+\tau\ln(A))\gamma^m}{(1-\gamma)^3}\left[\left(1+\tau\ln\left(A\,K^4\right)\right)\right.$$
$$\left(\left(\ln(A\,K^4)+\frac{1+\tau\ln(A)}{\tau(1-\gamma)}\right)\left(\frac{1}{t}+(\gamma^m)^{t/2}\right)+\frac{\sqrt{\tau}}{K}\right)$$
$$\left.+\,\tau\,(\ln(A)+1)\,\frac{1}{K^2}\right]+\frac{\tau}{(1-\gamma)^2}\ln(A)$$

- **SPMA***: if* $\eta_t = \eta = \min\left\{\frac{1-\gamma}{2}, \frac{(1-\gamma)\sqrt{\ln(A)}}{\sqrt{K}}\right\}$*, then, for all* $t \in [K]$*,*

$$\epsilon_t := \|\boldsymbol{\epsilon}_t\|_\infty \leq \frac{\sqrt{\ln(A)}\,\gamma^m}{(1-\gamma)^3}\,\frac{1}{\sqrt{K}}$$

Proof. Using Theorem 11 for soft SPMA and Corollary 5 for SPMA, and following the same proof as Corollary 8. Note that the requirement for $\eta \leq \frac{1-\gamma}{2}$ is to ensure $\ln(1+\eta_t\,\Delta^t(s,\cdot))$ in the SPMA update is well-defined (see the proof in Corollary 11 for details) □

B.3.3 Proof of Corollary 2

Generic Regret Bound

Theorem 12 (Generic Regret Bound)**.** *Consider a sequence of linear functions* $f_t(\pi) := \langle \pi, d_t\rangle$ *for a sequence of vectors* $\{d_0, d_1, \ldots, d_{K-1}\}$ *s.t.* $\|d_t\|_\infty \leq D_t$*. Consider the following update at iteration* $t \in [K]$*, if* η_t *is a step-size sequence,* $\tau_t = \eta_t\,\tau$*,* π_0 *is the uniform distribution and*

$$\pi_{t+1} = \arg\min_{\pi\in\Delta_A}\left\{\langle \pi, d_t\rangle + KL(\pi||\pi_t) + \mathcal{R}_t(\pi)\right\}$$
$$\mathcal{R}_t(\pi) := \tau_t\,\mathcal{R}(\pi)$$
$$\mathcal{R}(\pi) := \ln(A) - \mathcal{H}(\pi) \geq 0$$

then, for any comparator $u \in \Delta_A$*,*

$$\sum_{t=0}^{K-1}\left[\frac{\langle \pi_t - u,\, d_t\rangle}{\eta_t} + \tau\,[\mathcal{H}(u) - \mathcal{H}(\pi_t)]\right] \leq \sum_{t=0}^{K-1}\left[\frac{KL(u||\pi_t)}{\eta_t} - \frac{KL(u||\pi_{t+1})}{\eta_t} - \tau KL(u||\pi_{t+1})\right]$$
$$+\sum_{t=0}^{K-1}\frac{D_t^2}{2\,\eta_t}$$

Proof. The following properties will be helpful in proving the theorem. For policies π, π' and comparator u,

$$\mathcal{R}_t(\pi) - \mathcal{R}_t(\pi') = \tau_t \langle \ln(\pi), \pi - \pi' \rangle - \tau_t \, \mathrm{KL}(\pi' || \pi) \qquad \text{(Entropy property)}$$
$$\langle u - \pi', \ln(\pi') - \ln(\pi) \rangle = \mathrm{KL}(u||\pi) - \mathrm{KL}(u||\pi') - \mathrm{KL}(\pi'||\pi) \qquad \text{(3 point property)}$$
$$\langle \pi - \pi_{t+1}, d_t + \ln(\pi_{t+1}) - \ln(\pi_t) + \tau_t \ln(\pi_{t+1}) \rangle \geq 0 \qquad \text{(Optimality condition)}$$

$$\begin{aligned}
[f_t(\pi_t) - f_t(u)] + \mathcal{R}_t(\pi_{t+1}) - \mathcal{R}_t(u) &= \langle \pi_t - u,\, d_t \rangle + \langle \tau_t \ln(\pi_{t+1}),\, \pi_{t+1} - u \rangle \\
&\quad - \tau_t \mathrm{KL}(u||\pi_{t+1}) \\
&\qquad \text{(Entropy property with } \pi = \pi_{t+1},\, \pi' = u) \\
&= \langle \pi_{t+1} - u,\, d_t \rangle + \langle \pi_t - \pi_{t+1},\, d_t \rangle \\
&\quad + \langle \tau_t \ln(\pi_{t+1}),\, \pi_{t+1} - u \rangle - \tau_t \mathrm{KL}(u||\pi_{t+1}) \\
&= \underbrace{\langle \pi_{t+1} - u,\, d_t + \tau_t \ln(\pi_{t+1}) - \ln(\pi_t) + \ln(\pi_{t+1}) \rangle}_{\leq 0 \text{ by the optimality condition for } \pi = u} \\
&\quad + \langle \pi_{t+1} - u,\, \ln(\pi_t) - \ln(\pi_{t+1}) \rangle + \langle \pi_t - \pi_{t+1},\, d_t \rangle \\
&\quad - \tau_t \mathrm{KL}(u||\pi_{t+1}) \qquad \text{(Dropping the negative term)} \\
&\leq \langle \pi_{t+1} - u,\, \ln(\pi_t) - \ln(\pi_{t+1}) \rangle + \langle \pi_t - \pi_{t+1},\, d_t \rangle \\
&\quad - \tau_t \mathrm{KL}(u||\pi_{t+1}) \\
&= \mathrm{KL}(u||\pi_t) - \mathrm{KL}(u||\pi_{t+1}) - \mathrm{KL}(\pi_{t+1}||\pi_t) \\
&\quad + \langle \pi_t - \pi_{t+1},\, d_t \rangle - \tau_t \mathrm{KL}(u||\pi_{t+1}) \\
&\qquad \text{(3 point property with } u = u, \pi = \pi_t, \pi' = \pi_{t+1}) \\
&\leq \mathrm{KL}(u||\pi_t) - \mathrm{KL}(u||\pi_{t+1}) - \mathrm{KL}(\pi_{t+1}||\pi_t) \\
&\quad - \tau_t \mathrm{KL}(u||\pi_{t+1}) + \frac{1}{2}||\pi_t - \pi_{t+1}||_1^2 + \frac{1}{2}||d_t||_\infty^2 \\
&\qquad \text{(Fenchel-Young inequality)} \\
&\leq \mathrm{KL}(u||\pi_t) - \mathrm{KL}(u||\pi_{t+1}) - \mathrm{KL}(\pi_{t+1}||\pi_t) \\
&\quad - \tau_t \mathrm{KL}(u||\pi_{t+1}) + \mathrm{KL}(\pi_{t+1}||\pi_t) + \frac{1}{2}||d_t||_\infty^2 \\
&\qquad \text{(Pinsker's inequality)} \\
&= \mathrm{KL}(u||\pi_t) - \mathrm{KL}(u||\pi_{t+1}) - \tau_t \mathrm{KL}(u||\pi_{t+1}) + \frac{1}{2}||d_t||_\infty^2 \\
&\leq \mathrm{KL}(u||\pi_t) - \mathrm{KL}(u||\pi_{t+1}) - \tau_t \mathrm{KL}(u||\pi_{t+1}) + \frac{D_t^2}{2} \\
&\qquad \text{(Since } \|d_t\|_\infty \leq D_t) \\
&= \mathrm{KL}(u||\pi_t) - \mathrm{KL}(u||\pi_{t+1}) - \eta_t \tau \mathrm{KL}(u||\pi_{t+1}) + \frac{D_t^2}{2} \\
&\qquad \text{(Since } \tau_t = \eta_t \, \tau)
\end{aligned}$$

Rearranging and dividing both-sides by η_t we get

$$\frac{\langle \pi_t - u,\, d_t \rangle}{\eta_t} + \mathcal{R}(\pi_{t+1}) - \mathcal{R}(u) \leq \frac{\mathrm{KL}(u||\pi_t)}{\eta_t} - \frac{\mathrm{KL}(u||\pi_{t+1})}{\eta_t} - \tau \mathrm{KL}(u||\pi_{t+1}) + \frac{D_t^2}{2\,\eta_t}$$

$$\begin{aligned}\frac{\langle \pi_t - u,\, d_t \rangle}{\eta_t} + \mathcal{R}(\pi_t) - \mathcal{R}(u) &\leq [\mathcal{R}(\pi_t) - \mathcal{R}(\pi_{t+1})] + \frac{\mathrm{KL}(u||\pi_t)}{\eta_t} - \frac{\mathrm{KL}(u||\pi_{t+1})}{\eta_t} \\ &\quad - \tau \mathrm{KL}(u||\pi_{t+1}) + \frac{D_t^2}{2\,\eta_t}\end{aligned}$$

Summing from $t = 0$ to $K - 1$,

$$\begin{aligned}\sum_{t=0}^{K-1} \left[\frac{\langle \pi_t - u,\, d_t \rangle}{\eta_t} + \mathcal{R}(\pi_t) - \mathcal{R}(u)\right] &\leq [\mathcal{R}(\pi_0) - \mathcal{R}(\pi_K)] + \sum_{t=0}^{K-1} \left[\frac{\mathrm{KL}(u||\pi_t)}{\eta_t} - \frac{\mathrm{KL}(u||\pi_{t+1})}{\eta_t} \right. \\ &\quad \left. - \tau \mathrm{KL}(u||\pi_{t+1}) + \frac{D_t^2}{2\,\eta_t}\right] \\ &\leq \sum_{t=0}^{K-1} \left[\frac{\mathrm{KL}(u||\pi_t)}{\eta_t} - \frac{\mathrm{KL}(u||\pi_{t+1})}{\eta_t} - \tau \mathrm{KL}(u||\pi_{t+1})\right] \\ &\quad + \sum_{t=0}^{K-1} \frac{D_t^2}{2\,\eta_t} \\ &\quad \text{(Since } \mathcal{R}(\pi_0) = \ln(A) - \mathcal{H}(\pi_0) = 0 \text{ and } \mathcal{R}(\pi_K) \geq 0)\end{aligned}$$

□

Regret Bound for (soft) NPG

In this section, we instantiate the regret and policy evaluation bounds for (soft) NPG

Corollary 10 (Regret Bounds)**.** *Suppose $\pi_0(\cdot|s)$ is the uniform distribution over actions for each state s. For any sequence $\{q_\zeta^t\}_{t=0}^{K-1}$ satisfying $\left\|q_\zeta^t\right\|_\infty \leq H_\tau$, the regret for (soft)* NPG *can be bounded as:*

- ***Soft* NPG***: Setting $\eta_t = \frac{1}{c+\tau\,(t+1)}$ for a constant $c \geq 0$ to be determined later, guarantees that,*

$$\begin{aligned}&\max_s \left|\sum_{t=0}^{K-1} \left[\langle \pi_\tau^*(\cdot|s) - \pi_t(\cdot|s),\, q_\zeta^t(s,\cdot)\rangle + \tau\left[\mathcal{H}(\pi_\tau^*(\cdot|s)) - \mathcal{H}(\pi_t(\cdot|s))\right]\right]\right| \\ &\quad \leq \frac{H_\tau^2}{2\,\tau}\left[1 + \ln(K)\right] + (c+\tau)\,\ln(A)\end{aligned}$$

- **NPG***: Setting $\eta_t = \eta = \frac{\sqrt{2}\,(1-\gamma)\sqrt{\ln(A)}}{\sqrt{K}}$ guarantees that,*

$$\max_s \left|\sum_{t=0}^{K-1} \left[\langle \pi^*(\cdot|s) - \pi_t(\cdot|s),\, q_\zeta^t(s,\cdot)\rangle\right]\right| \leq \frac{\sqrt{2\ln(A)}\,\sqrt{K}}{1-\gamma}$$

Proof. First note that by using the mirror descent view of the (soft) NPG update [Xiao, 2022], it can be equivalently written as: for all $s \in \mathcal{S}$,

$$\pi_{t+1}(\cdot|s) = \arg\min_{\pi \in \Delta} \Big[-\eta_t \langle q_\zeta^t(s,\cdot), \pi(\cdot|s) \rangle + \mathrm{KL}(\pi(\cdot|s)||\pi_t(\cdot|s)) - \tau_t \mathcal{H}(\pi(\cdot|s)) \Big]$$

By comparing to the update in Theorem 12, we note that $d_t = -\eta_t \, q_\zeta^t(s,\cdot)$ and $\|d_t\|_\infty = \eta_t \left\| q_\zeta^t \right\|_\infty \leq \eta_t \, H_\tau$. If $\tau_t = \eta_t \tau$ and $\pi_0(\cdot|s)$ is a uniform distribution for each state s, we can instantiate Theorem 12 for each state s, and obtain the following regret bound for the comparator u.

$$\begin{aligned}
&\sum_{t=0}^{K-1} \Big[\langle u(\cdot|s) - \pi_t(\cdot|s),\, q_\zeta^t(s,\cdot) \rangle + \tau \left[\mathcal{H}(u(\cdot|s)) - \mathcal{H}(\pi_t(\cdot|s)) \right] \Big] \\
&\qquad \leq \sum_{t=0}^{K-1} \left[\frac{\mathrm{KL}(u(\cdot|s)||\pi_t(\cdot|s))}{\eta_t} - \frac{\mathrm{KL}(u(\cdot|s)||\pi_{t+1}(\cdot|s))}{\eta_t} - \tau \mathrm{KL}(u(\cdot|s)||\pi_{t+1}(\cdot|s)) \right] \\
&\qquad\quad + \frac{H_\tau^2}{2} \sum_{t=0}^{K-1} \eta_t
\end{aligned}$$

Now we consider two cases corresponding to NPG and its soft variant.

Soft NPG: Using that $\tau \neq 0$, setting $u = \pi_\tau^*$ and bounding the RHS in the above inequality,

$$\begin{aligned}
&\sum_{t=0}^{K-1} \Big[\langle \pi_\tau^*(\cdot|s) - \pi_t(\cdot|s),\, q_\zeta^t(s,\cdot) \rangle + \tau \left[\mathcal{H}(\pi_\tau^*(\cdot|s)) - \mathcal{H}(\pi_t(\cdot|s)) \right] \Big] \\
&\quad \leq \sum_{t=1}^{K-1} \mathrm{KL}(\pi_\tau^*(\cdot|s)||\pi_t(\cdot|s)) \left[\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} - \tau \right] + \frac{1}{\eta_0} \mathrm{KL}(\pi_\tau^*(\cdot|s)||\pi_0(\cdot|s)) + \frac{H_\tau^2}{2} \sum_{t=0}^{K-1} \eta_t \\
&\quad = \frac{H_\tau^2}{2} \sum_{t=0}^{K-1} \frac{1}{c + \tau\,(t+1)} + (c+\tau)\, \mathrm{KL}(\pi_\tau^*(\cdot|s)||\pi_0(\cdot|s)) \qquad \text{(Setting } \eta_t = \tfrac{1}{c+\tau\,(t+1)}\text{)} \\
&\quad \leq \frac{H_\tau^2}{2} \sum_{t=0}^{K-1} \frac{1}{\tau\,(t+1)} + (c+\tau)\, \ln(A) \quad \text{(Since } \pi_0(\cdot|s) \text{ is a uniform distribution for all } s\text{)} \\
&\quad \leq \frac{H_\tau^2}{2\,\tau} \left[1 + \ln(K) \right] + (c+\tau)\, \ln(A) \qquad \text{(Since } \textstyle\sum_{t=1}^{K} 1/t \leq 1 + \ln(K)\text{)}
\end{aligned}$$

Since the above bound holds for all s,

$$\begin{aligned}
&\max_s \left| \sum_{t=0}^{K-1} \left[\langle \pi_\tau^*(\cdot|s) - \pi_t(\cdot|s),\, q_\zeta^t(s,\cdot) \rangle + \tau \left[\mathcal{H}(\pi_\tau^*(\cdot|s)) - \mathcal{H}(\pi_t(\cdot|s)) \right] \right] \right| \\
&\qquad \leq \frac{H_\tau^2}{2\,\tau} \left[1 + \ln(K) \right] + (c+\tau)\, \ln(A)
\end{aligned}$$

NPG: Using $u = \pi^*$ and a constant step-size i.e. $\eta_t = \eta$ for all t, in which case the regret bound can be simplified as:

$$\begin{aligned}\sum_{t=0}^{K-1}\left[\langle \pi^*(\cdot|s) - \pi_t(\cdot|s),\, q_\zeta^t(s,\cdot)\rangle\right] &\le \frac{1}{\eta}\mathrm{KL}(\pi^*(\cdot|s)||\pi_0(\cdot|s) + \frac{\eta K}{2\,(1-\gamma)^2} \\ &\le \frac{\ln(A)}{\eta} + \frac{\eta K}{2\,(1-\gamma)^2} \\ &\qquad\qquad \text{(Since } \pi_0 \text{ is the uniform distribution)} \\ &\le \frac{\sqrt{2\ln(A)}\,\sqrt{K}}{1-\gamma} \qquad \text{(Setting } \eta = \tfrac{\sqrt{2}\,(1-\gamma)\sqrt{\ln(A)}}{\sqrt{K}}\text{)}\end{aligned}$$

Since the above bound holds for all s,

$$\max_s \left|\sum_{t=0}^{K-1}\left[\langle \pi^*(\cdot|s) - \pi_t(\cdot|s),\, q_\zeta^t(s,\cdot)\rangle\right]\right| \le \frac{\sqrt{2\ln(A)}\,\sqrt{K}}{1-\gamma}.$$

□

Regret Bound for (soft) SPMA

In this section, we instantiate the regret and policy evaluation bounds for (soft) SPMA.

Corollary 11 (Regret Bounds)**.** *Suppose $\pi_0(\cdot|s)$ is the uniform distribution over actions for each state s, and let $\eta_t = \frac{1}{c+\tau\,(t+1)}$ for some constant $c \ge 0$ to be determined later. For any sequence $\{q_\zeta^t\}_{t=0}^{K-1}$ satisfying $\left\|q_\zeta^t\right\|_\infty \le H_\tau$, the regret for (soft)* SPMA *can be bounded as:*

- ***Soft* SPMA***: Setting $\eta_t = \frac{1}{c+\tau\,(t+1)}$ for a constant $c \ge 2\,\max\{H_\tau, \zeta\ln(A)\}$, guarantees that,*

$$\begin{aligned}&\max_s \left|\sum_{t=0}^{K-1}\left[\langle \pi_\tau^*(\cdot|s) - \pi_t(\cdot|s),\, q_\zeta^t(s,\cdot)\rangle + \tau\left[\mathcal{H}(\pi_\tau^*(\cdot|s)) - \mathcal{H}(\pi_t(\cdot|s))\right]\right]\right| \\ &\qquad \le \frac{3\,H_\tau^2}{\tau}\left[1+\ln(K)\right] + (c+\tau)\,\ln(A)\end{aligned}$$

- **SPMA***: Setting $\eta_t = \eta = \min\left\{\frac{1-\gamma}{2}, \frac{\sqrt{2}\,(1-\gamma)\sqrt{\ln(A)}}{\sqrt{K}}\right\}$ guarantees that,*

$$\max_s \left|\sum_{t=0}^{K-1}\left[\langle \pi^*(\cdot|s) - \pi_t(\cdot|s),\, q_\zeta^t(s,\cdot)\rangle\right]\right| \le \frac{7\,\sqrt{\ln(A)}\,\sqrt{K}}{\sqrt{2}\,(1-\gamma)} + \frac{2}{1-\gamma}\,\ln(A).$$

Proof. For a fixed state $s \in \mathcal{S}$, first note that the (soft) SPMA update in Equation (3.5b) can be equivalently be written as follows: if $\Delta^t(s,a) := q^t_\zeta(s,a) - v^t_\zeta(s)$ for $d_t := -\ln(1+\eta_t\,\Delta^t(s,\cdot))$,

$$\begin{aligned}\pi_{t+1}(\cdot|s) &= \arg\min_{\pi\in\Delta}\Big[\mathrm{KL}(\pi(\cdot|s)||\pi_t(\cdot|s)\ \Big[1+\eta_t\ \Big(\Delta^t(s,\cdot)\Big)\Big]) - \tau_t\mathcal{H}(\pi(\cdot|s))\Big]\\ &= \arg\min_{\pi\in\Delta}\left[\langle d_t, \pi(\cdot|s)\rangle + \mathrm{KL}(\pi(\cdot|s)||\pi_t(\cdot|s)) - \tau_t\mathcal{H}(\pi(\cdot|s))\right],\end{aligned}$$

where we require that $1+\eta_t\Delta^t(s,\cdot) \geq 0$. Note that since $\left\|q^t_\zeta(s,\cdot)\right\|_\infty \leq H_\tau$, $\left\|\Delta^t(s,\cdot))\right\|_\infty \leq H_\tau$ and we require that $\eta_t \leq \frac{1}{2H_\tau}$. With this choice, $|\eta_t\,\Delta^t(s,a)| \leq \frac{1}{2}$. By comparing to the update in Theorem 12, we note that $d_t = -\ln(1+\eta_t\,\Delta^t(s,\cdot))$. Since $|\ln(1+x)| \leq 2|x|$ for all $x \geq -\frac{1}{2}$,

$$|-\ln(1+\eta_t\,\Delta^t(s,a))| \leq 2\,\eta_t\,|\Delta^t(s,a)| \leq 2\,\eta_t\,H_\tau \implies \|d_t\|_\infty \leq 2\,\eta_t\,H_\tau$$

Hence, $D_t = 2\,\eta_t\,H_\tau$. If $\tau_t = \eta_t\tau$ and $\pi_0(\cdot|s)$ is a uniform distribution for each state s, we can instantiate Theorem 12 for each state s, and obtain the following regret bound for the comparator u,

$$\begin{aligned}&\sum_{t=0}^{K-1}\left[\frac{\langle u(\cdot|s)-\pi_t(\cdot|s),\,\ln(1+\eta_t\,\Delta^t(s,\cdot))\rangle}{\eta_t} + \tau\left[\mathcal{H}(u(\cdot|s)) - \mathcal{H}(\pi_t(\cdot|s))\right]\right]\\ &\quad\leq \sum_{t=0}^{K-1}\left[\frac{KL(u(\cdot|s)||\pi_t(\cdot|s))}{\eta_t} - \frac{KL(u(\cdot|s)||\pi_{t+1}(\cdot|s))}{\eta_t} - \tau KL(u(\cdot|s)||\pi_{t+1}(\cdot|s))\right]\\ &\qquad + 2H_\tau^2\sum_{t=0}^{K-1}\eta_t\end{aligned}$$

In order to simplify the above expression, first note that,

$$\begin{aligned}\langle\pi_t(\cdot|s),\,\ln(1+\eta_t\,\Delta^t(s,\cdot))\rangle &\leq \ln(1-\eta_t\,\zeta\,\mathcal{H}(\pi_t(\cdot|s)))\\ \text{(Using Jensen's inequality and the fact } \textstyle\sum_a \pi_t(a|s)\Delta^t(s,a) &= -\zeta\,\mathcal{H}(\pi_t(\cdot|s))\end{aligned}$$

If η_t is chosen such that $\eta_t\,\zeta\,\mathcal{H}(\pi_t(\cdot|s)) \leq \frac{1}{2}$, and since $\mathcal{H}(\pi) \in [0, \ln(A)]$, it suffices to ensure $\eta_t \leq \frac{1}{2\zeta\,\ln(A)}$, and use the fact $\ln(1+x) \leq x$ for $x > -1$ to guarantee:

$$\langle\pi_t(\cdot|s),\,\ln(1+\eta_t\,\Delta^t(s,\cdot))\rangle \leq -\eta_t\,\zeta\,\mathcal{H}(\pi_t(\cdot|s)) \tag{B.6}$$

On the other hand, since choosing $\eta_t \leq \frac{1}{2H_\tau}$ ensures $\ln(1+\eta_t\Delta^t(s,\cdot))$ is well-defined,

$$\begin{aligned}
\langle u(\cdot|s),\, \ln(1+\eta_t\Delta^t(s,\cdot))\rangle &\geq [\langle u(\cdot|s),\, \eta_t\Delta^t(s,\cdot)\rangle - \langle u(\cdot|s),\, \eta_t^2[\Delta^t(s,\cdot)]^2\rangle] \\
&\qquad\qquad \text{(since } \ln(1+x) \geq x - x^2 \text{ for } x > -1/2) \\
&= \eta_t \left(\langle u(\cdot|s),\, q_\zeta^t(s,\cdot)\rangle - v_\zeta^t(s)\right) - \eta_t^2 \langle u(\cdot|s),\, [\Delta^t(s,\cdot)]^2\rangle \\
&= \eta_t \left(\langle u(\cdot|s),\, q_\zeta^t(s,\cdot)\rangle - \langle \pi_t,\, q_\zeta^t(s,\cdot)\rangle - \zeta\mathcal{H}(\pi_t(\cdot|s))\right) \\
&\quad - \eta_t^2\langle u(\cdot|s),\, [\Delta^t(s,\cdot)]^2\rangle \\
&\qquad\qquad \text{(since } v_\zeta^t(s) = \langle \pi_t(\cdot|s),\, q_\zeta^t\rangle + \zeta\mathcal{H}(\pi_t(\cdot|s))) \\
&\geq \eta_t \left(\langle u(\cdot|s),\, q_\zeta^t(s,\cdot)\rangle - \langle \pi_t(\cdot|s),\, q_\zeta^t(s,\cdot)\rangle - \zeta\mathcal{H}(\pi_t(\cdot|s))\right) - \eta_t^2 H_\tau^2 \\
&\qquad\qquad \text{(since } \left\|\Delta^t(s,\cdot))\right\|_\infty \leq H_\tau)
\end{aligned}$$

Combining the above inequalities with Equation (B.6),

$$\begin{aligned}
\frac{\langle u(\cdot|s) - \pi_t(\cdot|s),\, \ln(1+\eta_t\,\Delta^t(s,\cdot))\rangle}{\eta_t} &\geq \langle u(\cdot|s),\, q_\zeta^t(s,\cdot)\rangle - \langle \pi_t,\, q_\zeta^t(s,\cdot)\rangle - \eta_t\, H_\tau^2 \\
&= \langle u(\cdot|s) - \pi_t(\cdot|s), q_\zeta^t(s,\cdot)\rangle - \eta_t H_\tau^2
\end{aligned}$$

Using the above relation with the regret expression,

$$\begin{aligned}
&\sum_{t=0}^{K-1} \left[\langle u(\cdot|s) - \pi_t(\cdot|s), q_\zeta^t(s,\cdot)\rangle + \tau\left[\mathcal{H}(u(\cdot|s)) - \mathcal{H}(\pi_t(\cdot|s))\right]\right] \\
&\quad \leq \sum_{t=0}^{K-1} \left[\frac{KL(u(\cdot|s)||\pi_t(\cdot|s))}{\eta_t} - \frac{KL(u(\cdot|s)||\pi_{t+1}(\cdot|s))}{\eta_t} - \tau KL(u(\cdot|s)||\pi_{t+1}(\cdot|s))\right] \\
&\qquad + 3\, H_\tau^2 \sum_{t=0}^{K-1} \eta_t
\end{aligned}$$

Note that we require $\eta_t \leq \frac{1}{2H_\tau}$ and $\eta_t \leq \frac{1}{2\zeta\ln(A)}$ simultaneously for all t. Hence, it is sufficient to ensure that $\eta_t \leq \frac{1}{2\,\max\{H_\tau, \zeta\ln(A)\}}$ for all t, and hence require that $c \geq 2\,\max\{H_\tau, \zeta\ln(A)\}$.

Now we consider two cases corresponding to SPMA and its soft variant.

Soft SPMA: Using that $\tau \neq 0$, setting $u = \pi^*_\tau$ and bounding the RHS in the above inequality,

$$\begin{aligned}
&\sum_{t=0}^{K-1} \left[\langle \pi^*_\tau(\cdot|s) - \pi_t(\cdot|s),\, q^t_\zeta(s,\cdot)\rangle + \tau\left[\mathcal{H}(\pi^*_\tau(\cdot|s)) - \mathcal{H}(\pi_t(\cdot|s))\right]\right] \\
&\quad \leq \sum_{t=1}^{K-1} \mathrm{KL}(\pi^*_\tau(\cdot|s)||\pi_t(\cdot|s)) \left[\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} - \tau\right] + \frac{1}{\eta_0}\mathrm{KL}(\pi^*_\tau(\cdot|s)||\pi_0(\cdot|s)) \\
&\qquad + 3\,H_\tau^2 \sum_{t=0}^{K-1} \eta_t \\
&\quad = 3\,H_\tau^2 \sum_{t=0}^{K-1} \frac{1}{c + \tau\,(t+1)} + (c+\tau)\,\mathrm{KL}(\pi^*_\tau(\cdot|s)||\pi_0(\cdot|s)) \qquad (\text{Since } \eta_t = \tfrac{1}{c+\tau\,(t+1)}) \\
&\quad \leq 3\,H_\tau^2 \sum_{t=0}^{K-1} \frac{1}{c + \tau\,(t+1)} + (c+\tau)\,\ln(A) \\
&\qquad\qquad (\text{Since } \pi_0(\cdot|s) \text{ is a uniform distribution for all } s) \\
&\quad \leq \frac{3\,H_\tau^2}{\tau} \sum_{t=0}^{K-1} \frac{1}{t+1} + (c+\tau)\,\ln(A) \\
&\quad \leq \frac{3\,H_\tau^2}{\tau}\,[1 + \ln(K)] + (c+\tau)\,\ln(A) \qquad (\text{Since } \textstyle\sum_{t=1}^{K} 1/t \leq 1 + \ln(K))
\end{aligned}$$

Since the above bound holds for all s,

$$\begin{aligned}
&\max_s \left|\sum_{t=0}^{K-1} \left[\langle \pi^*_\tau(\cdot|s) - \pi_t(\cdot|s),\, q^t_\zeta(s,\cdot)\rangle + \tau\left[\mathcal{H}(\pi^*_\tau(\cdot|s)) - \mathcal{H}(\pi_t(\cdot|s))\right]\right]\right| \\
&\qquad \leq \frac{3\,H_\tau^2}{\tau}\,[1 + \ln(K)] + (c+\tau)\,\ln(A)
\end{aligned}$$

SPMA: Using $u = \pi^*$, $\tau = \zeta = 0$, and a constant step-size i.e. $\eta_t = \eta$ for all t, in which case the regret bound can be simplified as:

$$\begin{aligned}
\sum_{t=0}^{K-1} \left[\langle \pi^*(\cdot|s) - \pi_t(\cdot|s),\, q^t_\zeta(s,\cdot)\rangle\right] &\leq \frac{1}{\eta}\mathrm{KL}(\pi^*(\cdot|s)||\pi_0(\cdot|s)) + \frac{3\,\eta K}{(1-\gamma)^2} \\
&\leq \frac{\ln(A)}{\eta} + \frac{3\,\eta K}{(1-\gamma)^2} \\
&\qquad (\text{Since } \pi_0(\cdot|s) \text{ is a uniform distribution for all } s)
\end{aligned}$$

Recall that in the presence of entropy ensuring $\ln(1+\eta_t \Delta^t(s,\cdot))$ is well-defined required us to choose $\eta_t \leq \frac{1}{2\,H_\tau}$. When $\tau = 0$ and $\eta_t = \eta$, the condition simplifies to $\eta \leq \frac{1-\gamma}{2}$. Setting $\eta = \min\left\{\frac{\sqrt{2}\,(1-\gamma)\,\sqrt{\ln(A)}}{\sqrt{K}}, \frac{1-\gamma}{2}\right\}$ and using the fact that $\frac{1}{\min\{a,b\}} = \max\{1/a, 1/b\}$

$$
\begin{aligned}
&\leq \ln(A)\, \max\left\{\frac{\sqrt{K}}{\sqrt{2}\,(1-\gamma)\,\sqrt{\ln(A)}}, \frac{2}{1-\gamma}\right\} \\
&\quad + \frac{3\,K}{(1-\gamma)^2}\, \min\left\{\frac{\sqrt{2}\,(1-\gamma)\,\sqrt{\ln(A)}}{\sqrt{K}}, \frac{1-\gamma}{2}\right\} \\
&\leq \frac{7\,\sqrt{\ln(A)}\,\sqrt{K}}{\sqrt{2}\,(1-\gamma)} + \frac{2}{1-\gamma}\, \ln(A) \\
&\qquad\qquad \text{(Since } \max\{a,b\} \leq a+b,\ \min\{a,b\} \leq a\text{)}
\end{aligned}
$$

Since the above bound holds for all s,

$$
\max_s \left| \sum_{t=0}^{K-1} \Big[\langle \pi^*(\cdot|s) - \pi_t(\cdot|s),\, q_\zeta^t(s,\cdot)\rangle\Big] \right| \leq \frac{7\,\sqrt{\ln(A)}\,\sqrt{K}}{\sqrt{2}\,(1-\gamma)} + \frac{2}{1-\gamma}\, \ln(A).
$$

□

B.3.4 Proof of Theorem 5

Finally, we put everything together, and in the following two subsections, we prove theorems that quantify the performance of (soft) NPG and (soft) SPMA when using the hard or soft Bellman operator (i.e., $\zeta = \tau$ or $\zeta = 0$) as well as the case of no entropy regularization (i.e, $\tau = 0$ and $\zeta = 0$).

Putting everything together for soft NPG

Theorem 13 (Sub-optimality of soft NPG)**.** *Let π_τ^* denote the optimal entropy-regularized policy with value function v_τ^*. Consider the soft* NPG *update with step size $\eta_t = \frac{1}{c+\tau\,(t+1)}$, $c \geq \max\left\{\frac{8\,(1+\tau\,\ln(A))}{(1-\gamma)}, 32\,\tau\,\ln(A)\right\}$ and, $\delta(\tau,\zeta) := \frac{|\tau-\zeta|\,\ln(A)}{1-\gamma}$. Let $\pi_0(\cdot|s)$ be the uniform policy over actions for all $s \in \mathcal{S}$ and assume the policy evaluation step in Equation* (3.1)*.*

Then the resulting mixture policy $\bar{\pi}_K$ satisfies the following sub-optimality bound,

$$\left\|v_\tau^{\bar{\pi}_K} - v_\tau^*\right\|_\infty \leq \frac{1}{K\,(1-\gamma)}\left[\frac{(1+\tau\,\ln(A))^2}{2\,\tau\,(1-\gamma)^2}\,[1+\ln(K)] + (c+\tau)\,\ln(A)\right] + \frac{16(1+\tau\ln(A))\gamma^m}{(1-\gamma)^4\,K}\Bigg[\left(1+\tau\ln\left(A\,K^4\right)\right)\left(\left(\ln(A\,K^4) + \frac{1+\tau\ln(A)}{\tau(1-\gamma)}\right)\left(1+\ln(K)+\frac{1}{1-\sqrt{\gamma}}\right) + \sqrt{\tau}\right) + \tau\,(\ln(A)+1)\,\frac{1}{K}\Bigg] + \frac{2\,\delta(\tau,\zeta)}{(1-\gamma)^2}$$

Proof. Plugging the regret bound in Corollary 10 for soft NPG into the regret part of Theorem 4 immediately gives the first part of the upper-bound:

$$\frac{\|\mathrm{Regret}(K)\|_\infty}{K\,(1-\gamma)} \leq \frac{1}{K\,(1-\gamma)}\left[\frac{(1+\tau\,\ln(A))^2}{2\,\tau\,(1-\gamma)^2}\,[1+\ln(K)] + (c+\tau)\,\ln(A)\right]$$

Using the result from Theorem 11 and Corollary 6 to upper-bound the error part $E_K := \frac{2\sum_{k\in[K]}\|\boldsymbol{\epsilon}_k\|_\infty}{K\,(1-\gamma)}$in Theorem 4 we obtain:

$$E_K \leq \frac{16(1+\tau\ln(A))\gamma^m}{(1-\gamma)^4\,K}\Bigg[\left(1+\tau\ln\left(A\,K^4\right)\right)\left(\left(\ln(A\,K^4) + \frac{1+\tau\ln(A)}{\tau(1-\gamma)}\right)\left(1+\ln(K)+\frac{1}{1-\sqrt{\gamma}}\right) + \sqrt{\tau}\right) + \tau\,(\ln(A)+1)\,\frac{1}{K}\Bigg] + \frac{2\,\delta(\tau,\zeta)}{(1-\gamma)^2}$$

Where the above inequality can be obtained using the fact $\sum_{t=1}^K \frac{1}{t} \leq 1+\ln(K)$ (using integration) and $\sum_{t=1}^K(\gamma^m)^{t/2} \leq \frac{1}{1-\gamma^{m/2}} \leq \frac{1}{1-\sqrt{\gamma}}$

□

Corollary 12 (Sub-optimality of soft NPG with critic entropy)**.** *Let π_τ^* denote the optimal entropy-regularized policy with value function v_τ^*. Consider the soft* NPG *update with step size $\eta_t = \frac{1}{c+\tau\,(t+1)}$ and $c \geq \max\left\{\frac{8\,(1+\tau\,\ln(A))}{(1-\gamma)}, 32\,\tau\,\ln(A)\right\}$. Let $\pi_0(\cdot|s)$ be the uniform policy over actions for all $s \in \mathcal{S}$ and assume the policy evaluation step in Equation* (3.1) *with $\zeta = \tau$.*

Then the resulting mixture policy $\bar{\pi}_K$ satisfies the following sub-optimality bound,

$$\begin{aligned}\left\|v_\tau^{\bar{\pi}_K} - v_\tau^*\right\|_\infty \leq \frac{1}{K\,(1-\gamma)} & \left[\frac{(1+\tau\,\ln(A))^2}{2\,\tau\,(1-\gamma)^2}\,[1+\ln(K)] + (c+\tau)\,\ln(A)\right] \\ & + \frac{16(1+\tau\ln(A))\gamma^m}{(1-\gamma)^4\,K}\Bigg[\left(1+\tau\ln\left(A\,K^4\right)\right) \\ & \qquad \Bigg(\left(\ln(A\,K^4) + \frac{1+\tau\ln(A)}{\tau(1-\gamma)}\right)\left(1+\ln(K)+\frac{1}{1-\sqrt{\gamma}}\right) \\ & \qquad\quad + \sqrt{\tau}\Bigg) \\ & \qquad + \tau\,(\ln(A)+1)\,\frac{1}{K}\Bigg]\end{aligned}$$

Proof. Using Theorem 4 with Corollary 10 for soft NPG, Theorem 11, Corollary 6 and the result from Theorem 13. □

Corollary 13 (Sub-optimality of soft NPG without critic entropy)**.** *Let π_τ^* denote the optimal entropy-regularized policy with value function v_τ^*. Consider the soft* NPG *update with step size $\eta_t = \frac{1}{c+\tau\,(t+1)}$ and $c \geq \max\left\{\frac{8\,(1+\tau\,\ln(A))}{(1-\gamma)}, 32\,\tau\,\ln(A)\right\}$. Let $\pi_0(\cdot|s)$ be the uniform policy over actions for all $s \in \mathcal{S}$ and assume the policy evaluation step in Equation* (3.1) *with $\zeta = 0$. Then the resulting mixture policy $\bar{\pi}_K$ satisfies the following sub-optimality bound,*

$$\begin{aligned}\left\|v_\tau^{\bar{\pi}_K} - v_\tau^*\right\|_\infty \leq \frac{1}{K\,(1-\gamma)} & \left[\frac{(1+\tau\,\ln(A))^2}{2\,\tau\,(1-\gamma)^2}\,[1+\ln(K)] + (c+\tau)\,\ln(A)\right] \\ & + \frac{16(1+\tau\ln(A))\gamma^m}{(1-\gamma)^4\,K}\Bigg[\left(1+\tau\ln\left(A\,K^4\right)\right) \\ & \qquad \Bigg(\left(\ln(A\,K^4) + \frac{1+\tau\ln(A)}{\tau(1-\gamma)}\right)\left(1+\ln(K)+\frac{1}{1-\sqrt{\gamma}}\right) \\ & \qquad\quad + \sqrt{\tau}\Bigg) \\ & \qquad + \tau\,(\ln(A)+1)\,\frac{1}{K}\Bigg] + \frac{2\,\tau\,\ln(A)}{(1-\gamma)^3}\end{aligned}$$

Proof. Using Theorem 4 with Corollary 10 for soft NPG, Theorem 11, Corollary 7 and the result from Theorem 13. □

Theorem 14 (NPG + policy evaluation without entropy regularization)**.** *If π^* is the optimal policy whose value function is equal to v^*, the* NPG *update with $\eta_t = \eta = \frac{\sqrt{2}\,(1-\gamma)\sqrt{\ln(A)}}{\sqrt{K}}$, $\pi_0(\cdot|s)$ as the uniform initial policy for each $s \in \mathcal{S}$ with the policy evaluation procedure in Equation* (3.1) *with $\zeta = 0$ satisfies the following sub-optimality bound for the mixture*

policy $\bar{\pi}_K$,

$$\left\|v^{\bar{\pi}_K} - v^*\right\|_\infty \leq \frac{\sqrt{2\ln(A)}}{\sqrt{K}\,(1-\gamma)^2} + \frac{4\,\sqrt{\ln(A)}\,\gamma^m}{\sqrt{K}\,(1-\gamma)^4}$$

Proof. Using Corollary 3 with Corollary 10 for NPG and Corollary 7. □

Putting everything together for soft SPMA

Theorem 15 (Sub-optimality of soft SPMA)**.** *Let* π_τ^* *denote the optimal entropy-regularized policy with value function* v_τ^**. Consider the soft* SPMA *update with step size* $\eta_t = \frac{1}{c+\tau\,(t+1)}$*,* $c \geq \max\left\{\frac{4\,(1+\tau\,\ln(A))}{(1-\gamma)}, 32\,\tau\,\ln(A)\right\}$*, and,* $\delta(\tau,\zeta) := \frac{|\tau-\zeta|\,\ln(A)}{1-\gamma}$*. Let* $\pi_0(\cdot|s)$ *be the uniform policy over actions for all* $s \in \mathcal{S}$ *and assume the policy evaluation step in Equation* (3.1)*. Then the resulting mixture policy* $\bar{\pi}_K$ *satisfies the following sub-optimality bound,*

$$\begin{aligned}\left\|v_\tau^{\bar{\pi}_K} - v_\tau^*\right\|_\infty \leq \frac{1}{K\,(1-\gamma)}&\left[\frac{3\,(1+\tau\,\ln(A))^2}{2\,\tau\,(1-\gamma)^2}\,[1+\ln(K)] + (c+\tau)\,\ln(A)\right] \\ &+ \frac{16(1+\tau\ln(A))\gamma^m}{(1-\gamma)^4\,K}\Bigg[\left(1+\tau\ln\left(A\,K^4\right)\right) \\ &\qquad\left(\left(\ln(A\,K^4) + \frac{1+\tau\ln(A)}{\tau(1-\gamma)}\right)\left(1+\ln(K)+\frac{1}{1-\sqrt{\gamma}}\right)\right. \\ &\qquad\left.+\sqrt{\tau}\right) \\ &\qquad+\tau\,(\ln(A)+1)\,\frac{1}{K}\Bigg] + \frac{2\,\delta(\tau,\zeta)}{(1-\gamma)^2}\end{aligned}$$

Proof. Using Theorem 4 with Corollary 11 for soft SPMA, Theorem 11, Corollary 8 and the facts from the proof of Theorem 13. □

Corollary 14 (Sub-optimality of soft SPMA with critic entropy)**.** *Let* π_τ^* *denote the optimal entropy-regularized policy with value function* v_τ^**. Consider the soft* SPMA *update with step size* $\eta_t = \frac{1}{c+\tau\,(t+1)}$*,* $c \geq \max\left\{\frac{4\,(1+\tau\,\ln(A))}{(1-\gamma)}, 32\,\tau\,\ln(A)\right\}$*. Let* $\pi_0(\cdot|s)$ *be the uniform policy over actions for all* $s \in \mathcal{S}$ *and assume the policy evaluation step in Equation* (3.1) *with* $\zeta = \tau$*.*

Then the resulting mixture policy $\bar{\pi}_K$ satisfies the following sub-optimality bound,

$$\begin{aligned}\left\|v_\tau^{\bar{\pi}_K} - v_\tau^*\right\|_\infty \leq \frac{1}{K\,(1-\gamma)} & \left[\frac{3\,(1+\tau\,\ln(A))^2}{2\,\tau\,(1-\gamma)^2}\,[1+\ln(K)] + (c+\tau)\,\ln(A)\right] \\ & + \frac{16(1+\tau\ln(A))\gamma^m}{(1-\gamma)^4\,K}\Bigg[\left(1+\tau\ln\left(A\,K^4\right)\right) \\ & \qquad \Bigg(\left(\ln(A\,K^4) + \frac{1+\tau\ln(A)}{\tau(1-\gamma)}\right)\left(1+\ln(K)+\frac{1}{1-\sqrt{\gamma}}\right) \\ & \qquad\quad + \sqrt{\tau}\Bigg) \\ & \qquad + \tau\,(\ln(A)+1)\,\frac{1}{K}\Bigg]\end{aligned}$$

Proof. Using Theorem 4 with Corollary 11 for soft SPMA, Theorem 11, Corollary 8, Theorem 15 and the facts from the proof of Theorem 13. □

Corollary 15 (Sub-optimality of soft SPMA without critic entropy)**.** *Let π_τ^* denote the optimal entropy-regularized policy with value function v_τ^*. Consider the soft* SPMA *update with step size $\eta_t = \frac{1}{c+\tau\,(t+1)}$, $c \geq \max\left\{\frac{4\,(1+\tau\,\ln(A))}{(1-\gamma)}, 32\,\tau\,\ln(A)\right\}$. Let $\pi_0(\cdot|s)$ be the uniform policy over actions for all $s \in \mathcal{S}$ and assume the policy evaluation step in Equation* (3.1) *with $\zeta = 0$. Then the resulting mixture policy $\bar{\pi}_K$ satisfies the following sub-optimality bound,*

$$\begin{aligned}\left\|v_\tau^{\bar{\pi}_K} - v_\tau^*\right\|_\infty \leq \frac{1}{K\,(1-\gamma)} & \left[\frac{3\,(1+\tau\,\ln(A))^2}{2\,\tau\,(1-\gamma)^2}\,[1+\ln(K)] + (c+\tau)\,\ln(A)\right] \\ & + \frac{16(1+\tau\ln(A))\gamma^m}{(1-\gamma)^4\,K}\Bigg[\left(1+\tau\ln\left(A\,K^4\right)\right) \\ & \qquad \Bigg(\left(\ln(A\,K^4) + \frac{1+\tau\ln(A)}{\tau(1-\gamma)}\right)\left(1+\ln(K)+\frac{1}{1-\sqrt{\gamma}}\right) \\ & \qquad\quad + \sqrt{\tau}\Bigg) \\ & \qquad + \tau\,(\ln(A)+1)\,\frac{1}{K}\Bigg] + \frac{2\,\tau\,\ln(A)}{(1-\gamma)^3}\end{aligned}$$

Proof. Using Theorem 4 with Corollary 11 for soft SPMA, Theorem 11, Corollary 9, Theorem 15 and the facts from the proof of Theorem 13. □

Theorem 16 (SPMA + policy evaluation without entropy regularization)**.** *If π^* is the optimal policy whose value function is equal to v^*, the* SPMA *update in Equation* (3.4) *with $\eta_t = \eta = \min\left\{\frac{1-\gamma}{2}, \frac{\sqrt{2}\,(1-\gamma)\sqrt{\ln(A)}}{\sqrt{K}}\right\}$, $\pi_0(\cdot|s)$ as the uniform initial policy for each $s \in \mathcal{S}$ with the policy evaluation procedure in Equation* (3.1) *with $\zeta = 0$ satisfies the following*

sub-optimality bound for the mixture policy $\bar{\pi}_K$,

$$\left\|v^{\bar{\pi}_K} - v^*\right\|_\infty \leq \frac{1}{K\,(1-\gamma)}\left[\frac{7\,\sqrt{\ln(A)}\,\sqrt{K}}{\sqrt{2}\,(1-\gamma)} + \frac{2}{1-\gamma}\,\ln(A)\right] + \frac{2\,\sqrt{\ln(A)}\,\gamma^m}{\sqrt{K}\,(1-\gamma)^4}$$

Proof. Using Corollary 3 with Corollary 11 for SPMA and Corollary 9. □

B.4 Helper Lemmas

Lemma 7. *For any constant $C \in (0, 1/2)$, if P and Q are discrete distributions with support A, if $\|P - Q\|_1 \leq \frac{1}{2}$,then,*

$$|\mathcal{H}(Q) - \mathcal{H}(P)| \leq \|Q - P\|_1\,\ln\left(\frac{A}{C}\right) + \left(\frac{\ln(A-1)}{2} + \sqrt{2}\right)\sqrt{C}$$

Proof. By Cover [1999, Theorem 17.3.3], if $\|Q - P\|_1 \leq \frac{1}{2}$, then,

$$|\mathcal{H}(Q) - \mathcal{H}(P)| \leq \|Q - P\|_1\,\ln\left(\frac{A}{\|Q - P\|_1}\right) \tag{B.7}$$

Furthermore, using Sason [2013, Theorem 3], we also know that,

$$|\mathcal{H}(Q) - \mathcal{H}(P)| \leq \frac{\ln(A-1)}{2}\,\|Q - P\|_1 + h\left(\frac{1}{2}\,\|Q - P\|_1\right),$$

where h is a binary entropy, i.e. for $0 < x < 1$, $h(x) = -x\ln(x) - (1-x)\ln(1-x)$. Using the fact that $h(x) \leq 2\sqrt{x(1-x)}$,

$$\begin{aligned}
&\leq \frac{\ln(A-1)}{2}\,\|Q - P\|_1 + \sqrt{2}\sqrt{\|Q - P\|_1\left(1 - \frac{\|Q - P\|_1}{2}\right)} \\
&\leq \frac{\ln(A-1)}{2}\,\|Q - P\|_1 + \sqrt{2}\sqrt{\|Q - P\|_1}
\end{aligned}$$

Assuming $\|Q - P\|_1 \leq 2$

$$|\mathcal{H}(Q) - \mathcal{H}(P)| \leq \left(\frac{\ln(A-1)}{2} + \sqrt{2}\right)\sqrt{\|Q - P\|_1} \tag{B.8}$$

We will use each of these inequalities for different cases of $\|Q - P\|_1$, where $\|Q - P\|_1 \leq \frac{1}{2}$.

Case (1): For any constant $C \in (0, 1/2)$, if $\|Q - P\|_1 \geq C$, then, using Equation (B.7),

$$|\mathcal{H}(Q) - \mathcal{H}(P)| \leq \|Q - P\|_1\,\ln\left(\frac{A}{C}\right)$$

Case (2): On the other hand, if $\|Q - P\|_1 \leq C$, then we use Equation (B.8) to get that,

$$|\mathcal{H}(Q) - \mathcal{H}(P)| \leq \left(\frac{\ln(A-1)}{2} + \sqrt{2}\right) \sqrt{C}$$

Combining both cases, if $\|Q - P\|_1 \leq \frac{1}{2}$, then, for a constant $C \in (0, 1/2)$,

$$|\mathcal{H}(Q) - \mathcal{H}(P)| \leq \|Q - P\|_1 \ln\left(\frac{A}{C}\right) + \left(\frac{\ln(A-1)}{2} + \sqrt{2}\right) \sqrt{C}$$

□

Lemma 8. *For any constant $C \in (0, 1/2)$, and $\alpha = \eta\,\tau$ for $0 < \eta \leq 1$ to be determined, if P and Q are discrete distribution with support A such that $\|Q - P\|_1 \leq \frac{1}{2}$, then,*

$$\frac{1}{2}\|Q - P\|_1^2 \leq \alpha(\mathcal{H}(Q) - \mathcal{H}(P)) \implies \|Q - P\|_1 \leq 4\,\eta\,\tau \ln(\frac{A}{C}) + 2\,\sqrt{\tau} C^{\frac{1}{4}}$$

Proof. If the condition on the left-hand side of the lemma is true we have:

$$\begin{aligned}
\|Q - P\|_1^2 &\leq 2\,\eta\,\tau(\mathcal{H}(Q) - \mathcal{H}(P)) \\
&\leq 2\,\eta\,\tau\left(2\,\|Q - P\|_1 \ln(\frac{A}{C}) + 2\,\sqrt{C}\right) && \text{(Using the result from Lemma 7)} \\
&\leq 4\,\eta\,\tau\,\|Q - P\|_1 \ln(\frac{A}{C}) + 4\,\tau\,\sqrt{C} && \text{(since } \eta \leq 1\text{)}
\end{aligned}$$

After completing the square w.r.t $\|Q - P\|_1$ we obtain:

$$\|Q - P\|_1 \leq 4\,\eta\,\tau\,\ln(\frac{A}{C}) + 2\,\sqrt{\tau}\,C^{\frac{1}{4}}$$

□

Lemma 9. *For all $k \in [K]$,*

$$\sum_{i=1}^{k} \frac{\gamma^{k-i}}{(i+1)} \leq \frac{2}{k\,(1-\gamma)} + \frac{\gamma^{k/2}}{1-\gamma}$$

Proof. Define $j = k - i$, in which case, we need to bound,

$$\sum_{j=0}^{k-1} \frac{\gamma^j}{k-j+1} = \underbrace{\sum_{j=0}^{\lfloor k/2 \rfloor} \frac{\gamma^j}{k-j+1}}_{\text{Term (i)}} + \underbrace{\sum_{j=\lfloor k/2 \rfloor + 1}^{k-1} \frac{\gamma^j}{k-j+1}}_{\text{Term (ii)}}$$

For bounding Term (i), note that $j \leq \frac{k}{2} + 1$, meaning that $k - j + 1 \geq \frac{k}{2}$. Hence,

$$\text{Term (i)} = \sum_{j=0}^{\lfloor k/2 \rfloor} \frac{\gamma^j}{k-j+1} \leq \frac{2}{k} \sum_{j=0}^{\lfloor k/2 \rfloor} \gamma^j \leq \frac{2}{k} \sum_{j=0}^{\infty} \gamma^j \leq \frac{2}{k} \frac{1}{1-\gamma}$$

For bounding Term (ii), note that since $j \leq k$, $k - j + 1 \geq 1$. Hence,

$$\text{Term (ii)} = \sum_{j=\lfloor k/2 \rfloor + 1}^{k-1} \frac{\gamma^j}{k-j+1} \leq \sum_{j=\lfloor k/2 \rfloor + 1}^{k-1} \gamma^j \leq \sum_{j=\lfloor k/2 \rfloor + 1}^{\infty} \gamma^j \leq \frac{\gamma^{k/2}}{1-\gamma}$$

Combining both terms,

$$\sum_{j=0}^{k-1} \frac{\gamma^j}{k-j+1} \leq \frac{2}{k} \frac{1}{1-\gamma} + \frac{\gamma^{k/2}}{1-\gamma}$$

□

Lemma 10. *For any state-action value function q, for any $m \geq 1$,*

$$|(T_\tau^{\pi_1} q)^m(s,a) - (T_\zeta^{\pi_1} q)^m(s,a)| \leq \frac{|\tau - \zeta|}{1-\gamma} \ln(A)$$

Proof.

$$(T_\tau^{\pi_1} q)^m(s,a) = \mathbb{E}_{\substack{s_{t+1} \sim \mathcal{P}(\cdot|s_t,a_t) \\ a_t \sim \pi_1(\cdot|s_t)}} \Bigg[\sum_{t=0}^{m-1} \gamma^t (r(s_t,a_t) - \tau \ln \pi_1(a_t|s_t)) \tag{B.9}$$

$$+ \gamma^m q(s_m,a_m) \,\Big|\, s_0 = s,\, a_0 = a \Bigg] \qquad \text{(By definition of } T_\tau^{\pi_1})$$

$$= \mathbb{E}\left[\sum_{t=0}^{m-1} \gamma^t (r(s_t,a_t)) + \gamma^m q(s_m,a_m)\right] - \tau\, \mathbb{E}\left[\sum_{t=0}^{m-1} \gamma^t \ln(\pi_1(a_t|s_t))\right] \tag{B.10}$$

$$+ \zeta\, \mathbb{E}\left[\sum_{t=0}^{m-1} \gamma^t \ln(\pi_1(a_t|s_t))\right] - \zeta\, \mathbb{E}\left[\sum_{t=0}^{m-1} \gamma^t \ln(\pi_1(a_t|s_t))\right]$$

$$\text{(Add/Subtract } \zeta\, \mathbb{E}\left[\textstyle\sum_{t=0}^{m-1} \gamma^t \ln(\pi_1(a_t|s_t))\right])$$

$$= (T_\zeta^{\pi_1} q)^m(s,a) - (\tau - \zeta)\, \mathbb{E}\left[\sum_{t=0}^{m-1} \gamma^t \ln(\pi_1(a_t|s_t))\right] \qquad \text{(By definition of } T_\zeta^{\pi_1})$$

$$= (T_\zeta^{\pi_1} q)^m(s,a) + (\tau - \zeta)\, \mathbb{E}\left[\sum_{t=0}^{m-1} \gamma^t\, \mathcal{H}(\pi_1(.|s_t))\right]$$

$$\text{(By definition of } \mathcal{H}(\pi_1(.|s_t)))$$

$$\leq (T_\zeta^{\pi_1} q)^m(s,a) + \frac{|\tau - \zeta|}{1-\gamma} \ln(A) \qquad \text{(Since } \mathbb{E}[\mathcal{H}(\pi_1(.|s_t))] \leq \ln(A))$$

□

B.5 Experimental Details and Additional Results

B.5.1 Details for Stable Baselines Experiments

Following Tomar et al. [2020], Asad et al. [2025b] we use the default hyperparameters from `stable-baselines3` [Raffin et al., 2021] for each method. This choice is motivated by prior work, which focuses on evaluating the effectiveness of different surrogate losses rather than performing exhaustive hyperparameter searches. Such searches are particularly impractical for CNN-based actor and critic networks, where tuning multiple hyperparameters (e.g., framestack, buffer size, minibatch size, discount factor) is computationally expensive. The full list of hyperparameters for the Atari experiments is provided in Table B.1.

For our forward and reverse KL objectives, we set η_t to a constant chosen via a small grid search over $[0.01, 0.1, 1.0]$. Following prior work [Haarnoja et al., 2018a, Christodoulou, 2019, Mnih et al., 2013] for policy evaluation, q_ζ^{t-1} in Equation (3.2) is parameterized using a separate target network, whose parameters are updated via an exponential moving average of those in the critic model. For on-policy `PPO`, we adopt the optimal hyperparameters reported in RL Baselines3 Zoo [Raffin, 2020]. For all experiments, we use $n = 1$ actor gradient updates unless otherwise stated.

Regarding computational resources, each Atari run was executed on Digital Research Alliance of Canada's Narval cluster using one GPU with 12 CPU cores and 64GB of memory; runs with $n = 1$ actor-gradient update took approximately 12 hours, while runs with $n = 10$ actor-gradient updates took approximately 32 hours.

Hyperparameter	**FKL Objectives**	**RKL Objectives**	**DSAC**	**DQN**
Reward normalization	✗	✗	✗	✗
Observation normalization	✗	✗	✗	✗
Orthogonal weight initialization	✗	✗	✗	✗
Value function clipping	✗	✗	✗	✗
Gradient clipping	✗	✗	✗	✗
Probability ratio clipping	✗	✗	✗	✗
Entropy coefficient	auto	auto	auto	ϵ-greedy
Adam step-size	3×10^{-4}			
Buffer size	10^6			
Minibatch size	256			
Framestack	4			
Number of environment copies	8			
Discount factor	0.99			
Total number of timesteps	10^7			
Number of runs for plot averages	5			
Confidence interval for plot runs	$\sim 95\%$			

Table B.1: Hyper-parameters for Atari and Retro experiments.

B.5.2 Comparison to GreedyAC

We examine the empirical finding in our DSAC ablation in Section 3.5 (Figure 3.2) on discrete classic control tasks. The results in Figures B.1 and B.2 support our observation: DSAC with a fixed, per-environment tuned entropy coefficient, with or without critic entropy, attains strong performance on par with the prior discrete actor-critic framework GreedyAC [Neumann et al., 2018].

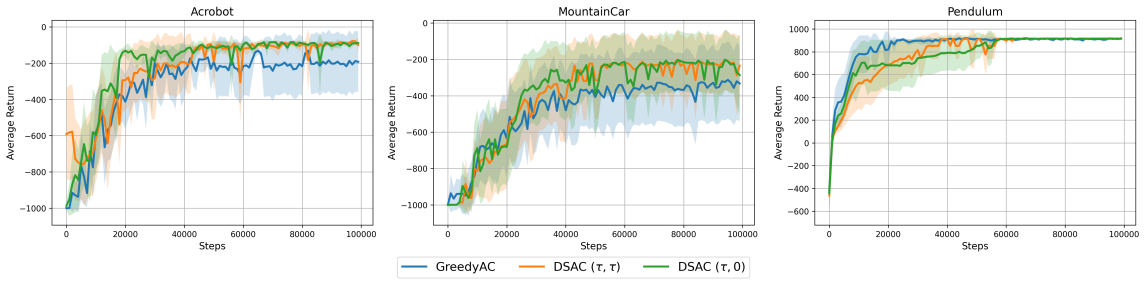


Figure B.1: Comparing GreedyAC with the default DSAC (with and without critic entropy and a fixed entropy coefficient) using batch size 32.

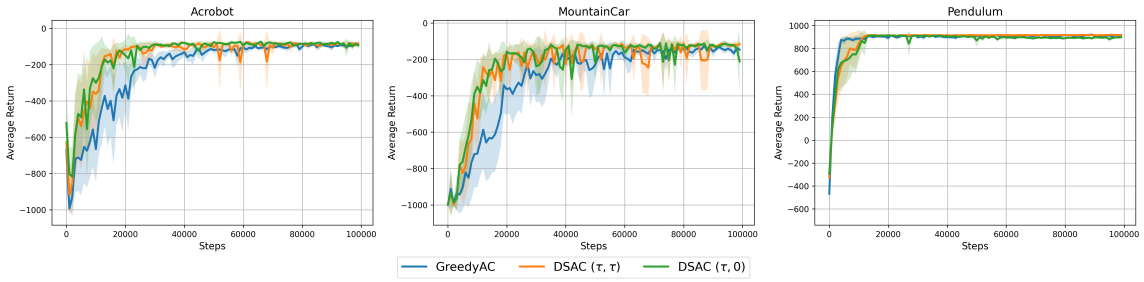


Figure B.2: Comparing GreedyAC with the default DSAC (with and without critic entropy and a fixed entropy coefficient) using batch size 256.

B.5.3 Does Double Q Learning Hurt or Improve Performance?

Zhou et al. [2022] hypothesize that DSAC suffers from two main failure modes. First, the coupled learning of the actor and critic can induce fluctuations in the Q-function distribution and policy entropy, leading to a sharp policy and poor performance. Second, taking the minimum of two Q functions in double Q-learning can induce Q-value underestimation and pessimistic exploration. To address these issues, they propose SD-SAC, which adds an entropy-penalty regularizer to prevent the policy entropy from deviating too much from its value at the previous iterate, together with clipped double-average Q-learning to stabilize the critic learning. Crucially, the reported SD-SAC experiments remove the automatic entropy tuning of the original DSAC and require tuning a fixed entropy coefficient. On the other hand, our work identifies critic entropy as the main bottleneck limiting DSAC's efficacy and proposes DSAC$(\tau, 0)$, which substantially outperforms the default DSAC while retaining the automatic entropy tuning of the original method.

The results in Table B.2 show that, unlike the findings of Zhou et al. [2022] for default DSAC, taking the minimum of two Q functions does not harm DSAC$(\tau, 0)$, even without Q clipping. Furthermore, both the single-Q and double-Q versions of DSAC$(\tau, 0)$ outperform SD-SAC.

Table B.2: IQM summary of final human-normalized scores with 95% stratified bootstrap confidence intervals: prior work hypothesizes that taking the minimum of two Q-networks harms DSAC and proposes SD-SAC, which clips and averages the two Q-values and uses an entropy penalty to prevent the collapse of policy entropy. However, the results show that the performance of our DSAC$(\tau, 0)$ with adaptive actor entropy does not degrade when taking the minimum of two Q-networks, and that both its single-Q and double-Q variants substantially outperform SD-SAC without clipping.

Method	IQM	CI Low	CI High
DSAC Double-Q $(\tau, 0)$	**0.9487**	0.8184	1.0789
DSAC $(\tau, 0)$	0.9226	0.8228	1.0288
SD-SAC (Baseline)	0.7605	0.6728	0.8545

Per-game comparison with SD-SAC. Figure B.3 complements the aggregate IQM results in Table B.2 by showing the final score on each Atari game. The broad per-game advantage of both DSAC$(\tau, 0)$ variants supports the paper's conclusion that critic entropy, rather than double-Q learning itself, is the principal limitation in the default discrete SAC formulation.

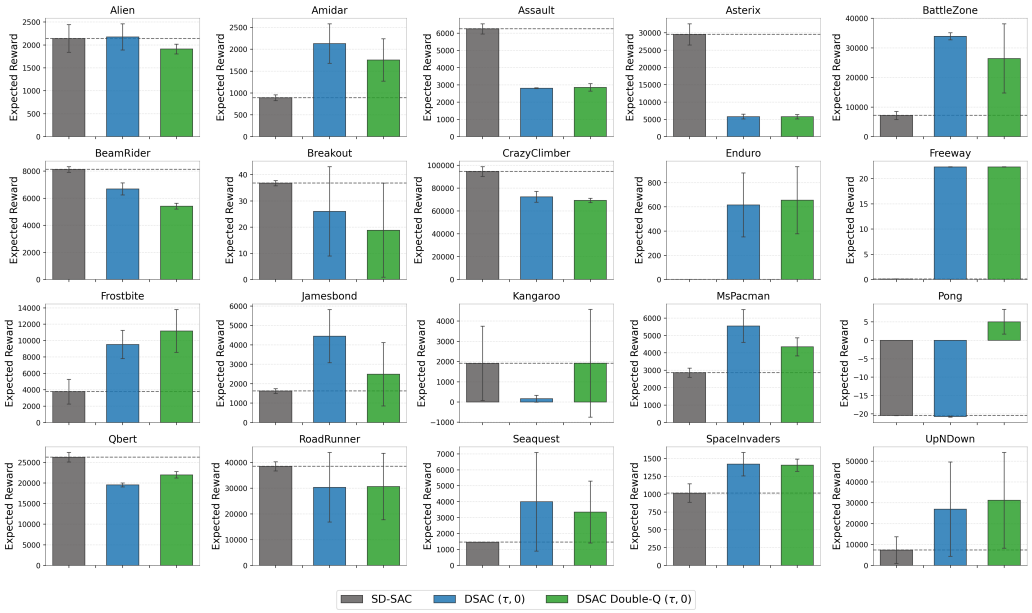


Figure B.3: Expected reward after 10M steps on each of the 20 Atari games, comparing our DSAC$(\tau, 0)$ and DSAC Double-Q $(\tau, 0)$ against SD-SAC. While the table above summarizes aggregate performance across games using IQM, this figure shows the per-game final rewards.

B.5.4 Is Fixed Target Entropy the Main Bottleneck in DSAC?

Table B.3: IQM summary of final human-normalized scores with 95% stratified bootstrap confidence intervals: prior work [Xu et al., 2021] hypothesizes that using a fixed target entropy in the adaptive entropy-coefficient scheme of Christodoulou [2019] harms DSAC, and proposes TES-SAC, which instead uses an adaptive target entropy. The results show that our DSAC$(\tau, 0)$, despite using a fixed target entropy, substantially outperforms TES-SAC.

Method	IQM	CI Low	CI High
DSAC $(\tau, 0)$	**0.9226**	0.8228	1.0288
TES-SAC (Baseline)	0.7210	0.6390	0.7965

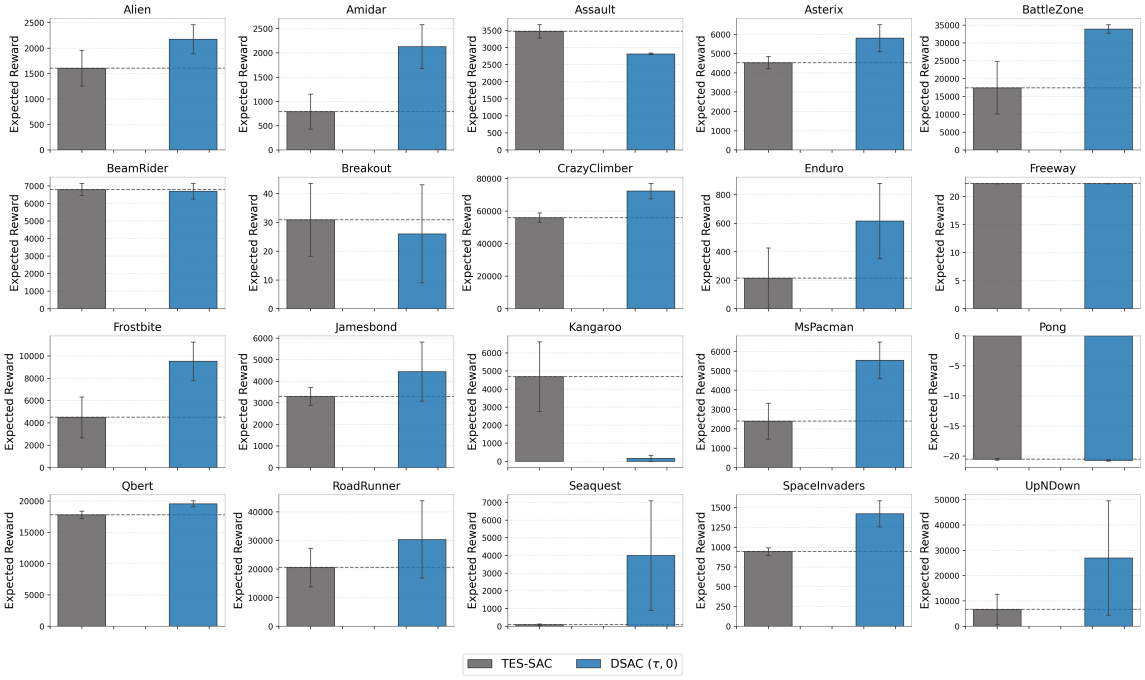


Figure B.4: Expected reward after 10M steps on each of the 20 Atari games, comparing our DSAC$(\tau, 0)$ against TES-SAC. While the table above summarizes aggregate performance across games using IQM, this figure shows the per-game final rewards.

B.5.5 Evaluating our DSAC Variant on Hard Exploration Games

Table 1 in [Taiga et al., 2021] provides a taxonomy of Atari games and identifies hard exploration games, further dividing them into dense- and sparse-reward settings. We compare our DSAC$(\tau, 0)$ with adaptive entropy against the baselines on 10 hard exploration games, comprising 5 dense-reward games (Alien, Amidar, Frostbite, MsPacman, Qbert) and 5 sparse-reward games (Freeway, Gravitar, Pitfall, Solaris, Venture). The results in Table B.4 show that, when evaluation is restricted to these hard exploration games, DSAC$(\tau, 0)$ substantially outperforms the baselines. The per-game final expected rewards are provided in Figure B.5.

Table B.4: IQM summary of final human-normalized scores with 95% stratified bootstrap confidence intervals: comparing DSAC$(\tau, 0)$, which uses adaptive actor entropy and disables critic entropy, against baselines on hard-exploration games with dense and sparse reward structures. Our DSAC variant achieves significantly higher overall performance than the baselines.

Method	IQM	CI Low	CI High
DSAC $(\tau, 0)$	**0.6951**	0.6497	0.7413
Adam LMCDQN (Baseline)	0.5862	0.5470	0.6299
DQN (Baseline)	0.4742	0.4558	0.4958
PPO (Baseline)	0.3225	0.2664	0.3825
TES-SAC (Baseline)	0.3022	0.2497	0.3485
SD-SAC (Baseline)	0.2053	0.1928	0.2188

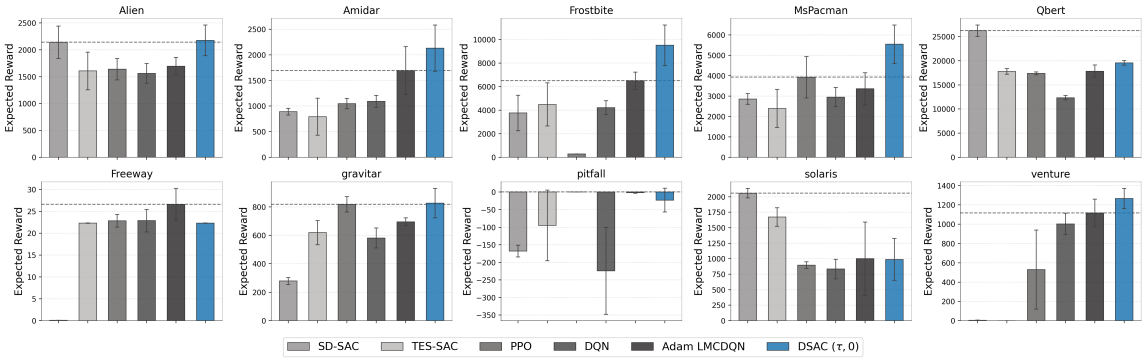


Figure B.5: Expected reward after 10M steps on each of the 10 hard exploration games, comparing our DSAC variant with adaptive actor entropy and no critic entropy against baselines (grey). The table above summarizes aggregate performance using IQM, while this figure shows per-game final performance.

B.5.6 Evaluating our DSAC Variant on stable-retro games

In addition to the standard Atari benchmark, we test our hypothesis on the role of critic entropy in DSAC using stable-retro games [Poliquin, 2026] with diverse reward structures and exploration levels. Importantly, stable-retro has a substantially larger action space than Atari, ranging from 18–468 actions compared to 4–18. The results in Figure B.6 show that our DSAC variant, which disables critic entropy while retaining automatic entropy tuning for the actor, yields substantial performance gains.

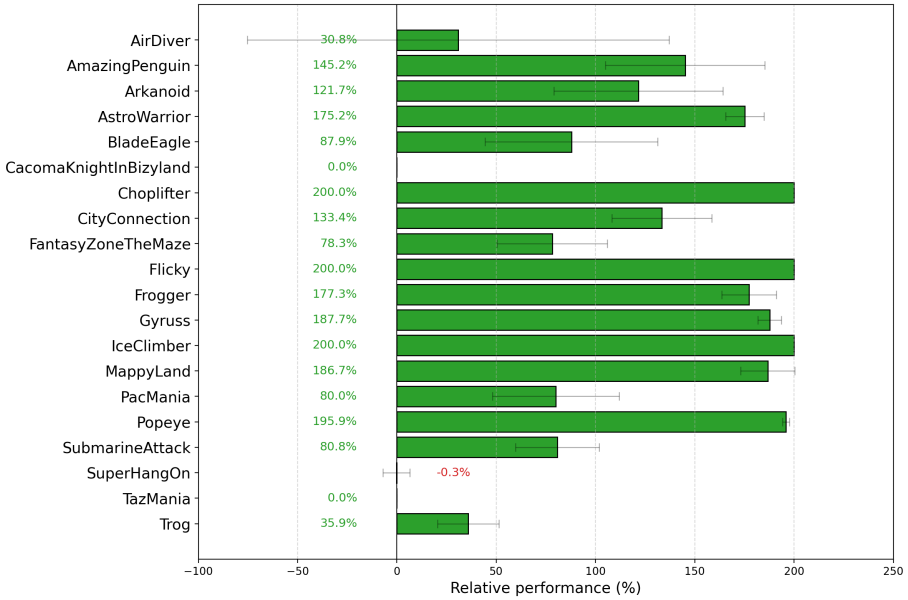


Figure B.6: Relative performance of DSAC$(\tau, 0)$, which uses adaptive actor entropy and disables critic entropy (our contribution), compared to the default DSAC across 20 stable-retro games. Results are averaged over 5 seeds with 95% confidence intervals. Consistent with the Atari results, removing critic entropy yields substantial performance gains.

B.5.7 Our Objectives Benefit Similarly from the Hard Bellman Operator

When using the forward and reverse KL-based objectives in our framework, we consistently observe the same trend as with DSAC: As observed in Figure B.7, employing the hard Bellman operator ($\zeta = 0$) with an adaptive entropy coefficient leads to substantial performance improvement compared to using the soft Bellman operator ($\zeta = \tau$).

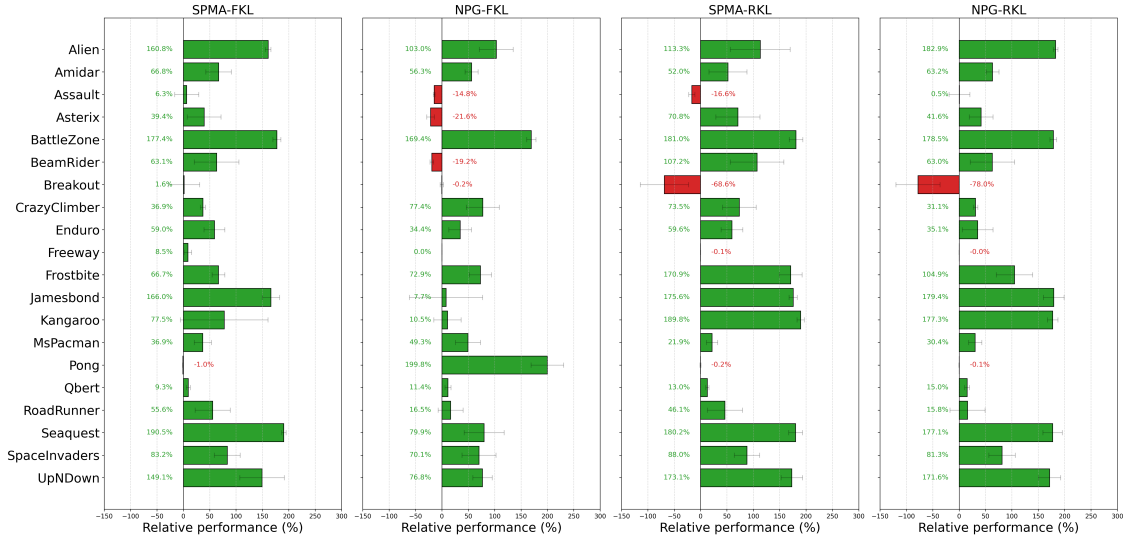


Figure B.7: Relative performance of forward and reverse KL-based objectives when critic entropy is disabled, measured against the same objectives with critic entropy enabled. Disabling critic entropy while retaining the adaptive entropy coefficient loss, as in DSAC, leads to performance improvements for all objectives.

While choosing $\zeta = 0$ is practical and avoids per-game hyperparameter search, we also consider an intermediate variant of our DSAC in which ζ is a nonzero constant selected from $\{0.001, 0.01, 0.1, 0.5\}$. The results in Table B.5 indicate that tuning a constant nonzero ζ can improve performance, but only at the cost of per-game hyperparameter search.

Table B.5: IQM summary of final human-normalized scores with 95% stratified bootstrap confidence intervals: comparing DSAC$(\tau, 0)$ (adaptive actor entropy and critic entropy disabled) vs. DSAC$(\tau, \text{const-}\zeta)$ (adaptive actor entropy and constant non-zero critic entropy).

Method	IQM	CI Low	CI High
DSAC $(\tau, \text{const-}\zeta)$	**0.9495**	0.8318	1.0707
DSAC $(\tau, 0)$	0.9226	0.8228	1.0288

Distribution of the tuned critic-entropy coefficient. Figure B.8 shows that the best nonzero value of ζ varies across games, with the largest concentration at 0.01 and 0.001. This variation reinforces the paper's recommendation to use $\zeta = 0$: it removes critic entropy and avoids a costly per-game hyperparameter search while retaining strong aggregate performance.

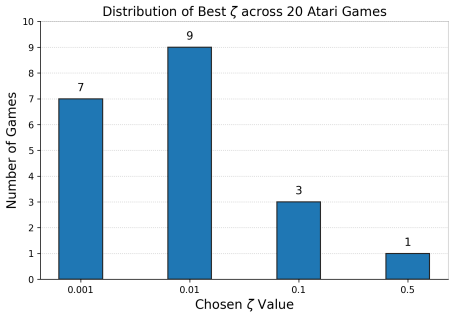


Figure B.8: Distribution of the best nonzero constant ζ across games for our DSAC$(\tau, \text{const-}\zeta)$ variant.

B.5.8 Additional Results: Is Entropy Regularization Necessary?

Effect of additional actor optimization without entropy. Figures B.9 and B.10 explain why the proposed objectives remain effective when explicit entropy regularization is removed. Increasing the number of actor updates from $n = 1$ to $n = 10$ stabilizes policy entropy and substantially improves performance, whereas DSAC does not obtain the same benefit.

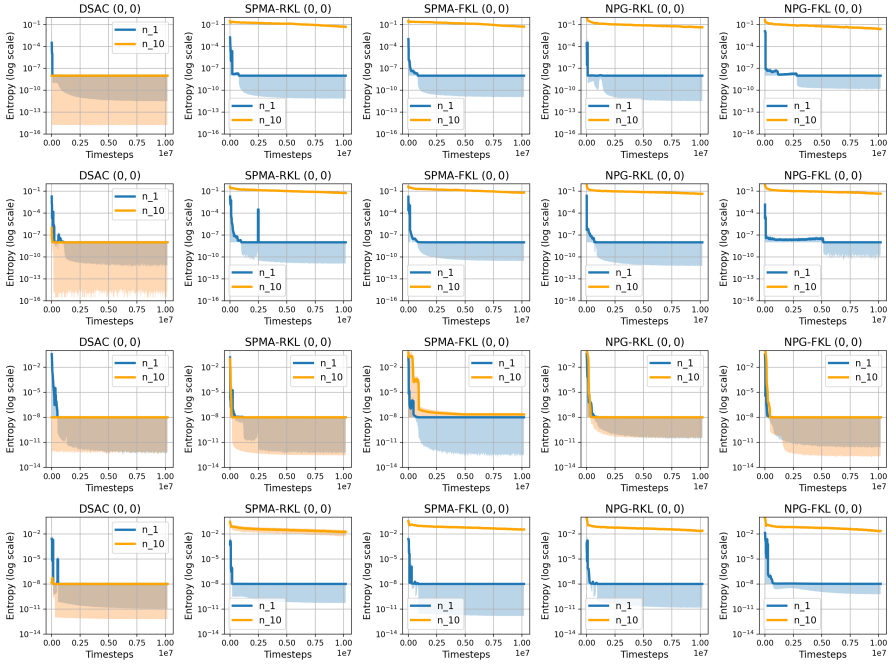


Figure B.9: **Policy Shannon entropy during training**. Without entropy regularization, increasing n from 1 to 10 substantially increases and stabilizes the Shannon entropy for our forward and reverse KL-based methods, compared to DSAC, across four Atari games. The rows correspond to MsPacman-v4, Seaquest-v4, Freeway-v4, and Breakout-v4.



Figure B.10: Effect of the number of actor optimization steps n without entropy regularization. Unlike DSAC, the performance of all objectives derived from our framework significantly boosts when n is increased from 1 to 10.

KL-divergence diagnostics. Figure B.11 provides a complementary view of the same mechanism. When entropy regularization is removed and n is increased to 10, our proposed forward- and reverse-KL actors maintain higher and more stable KL divergence than DSAC. This result supports the paper's broader finding that the proposed KL-based objectives are robust to removing explicit entropy bonuses, whereas DSAC, in the absence of entropy regularization, has no remaining source of regularization and KL collapses.

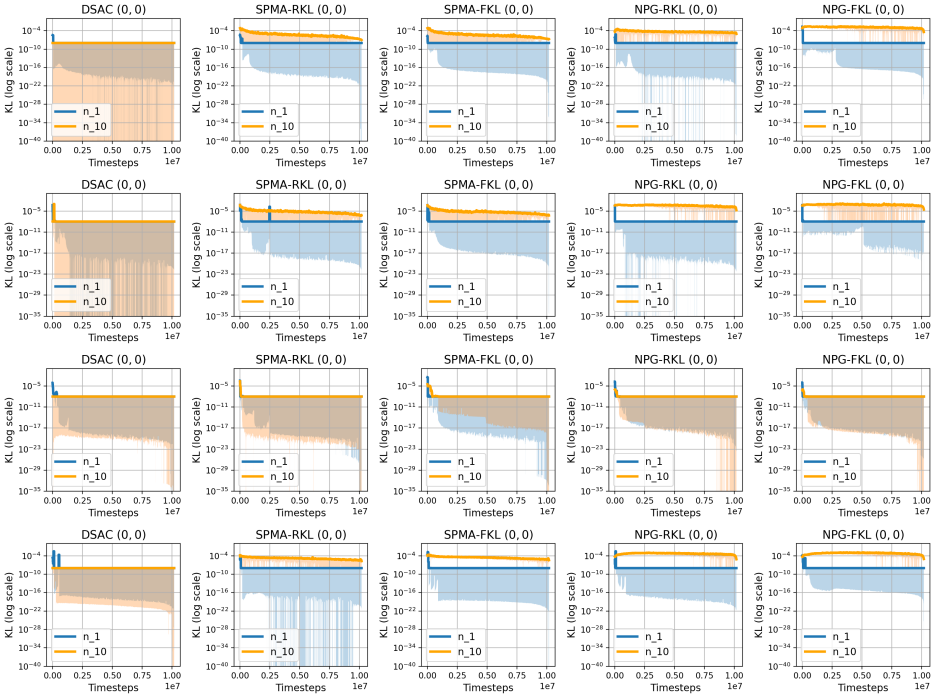


Figure B.11: **KL divergence during training.** Consistent with the Shannon entropy results, in the absence of entropy regularization, increasing n to 10 allows our forward- and reverse-KL-based actors to achieve higher KL divergence than DSAC, which also stabilizes over the course of training. The rows correspond to MsPacman-v4, Seaquest-v4, Freeway-v4, and Breakout-v4.

Performance without entropy regularization. Figure B.12 translates the preceding diagnostics into final performance. With sufficient actor optimization, the SPMA- and NPG-based objectives remain competitive even when both actor and critic entropy are disabled, directly supporting the paper's claim that their robustness is not solely produced by entropy regularization.

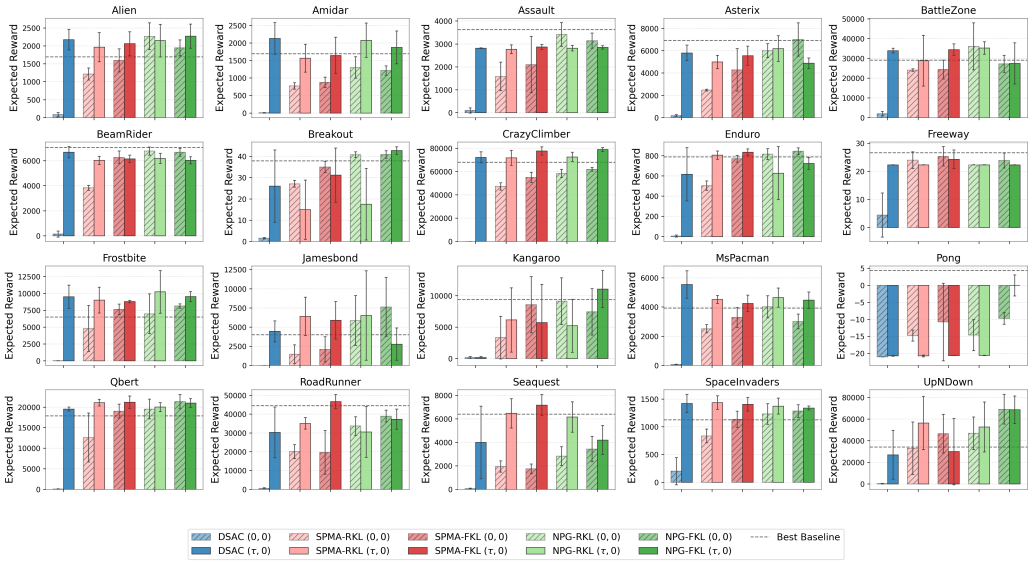


Figure B.12: **Entropy regularization ablation: disabling entropy.** Expected reward after 10M steps. Unlike `DSAC`, objectives from our framework, including the novel `SPMA-RKL` and `NPG-FKL`, remain robust with both actor and critic entropy disabled (dashed bars); solid bars denote the actor-entropy-enabled setting ($\tau \neq 0$). The dashed line denotes the best baseline among `PPO`, `DQN`, and `Adam LMCDQN`.

B.5.9 Expected reward after 10M steps

Disaggregated final performance. Figure B.13 complements the aggregate IQM comparison in Table 3.3 with per-game final rewards. The gains are distributed across the benchmark rather than being driven by a small number of games, strengthening the paper's conclusion that disabling critic entropy and using the proposed objectives produces broadly competitive Atari performance.

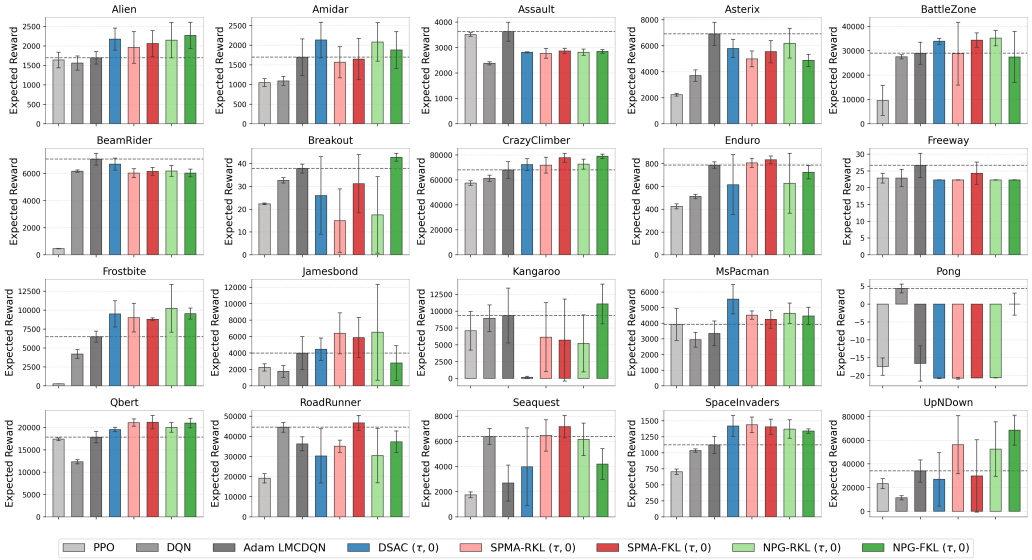


Figure B.13: **Expected reward after 10M steps**. While Table 3.3 reports aggregate performance using IQM, this figure shows the corresponding per-game final rewards. Objectives from our proposed framework (colored) use adaptive actor entropy ($\tau \neq 0$) and no critic entropy ($\zeta = 0$); baselines are shown in grey.

B.5.10 Full Horizon Training Curves Across 20 Games

Comparison with PPO. Figure B.14 shows the full training trajectories against the on-policy PPO baseline. The proposed off-policy objectives generally learn faster and reach higher returns, providing a temporal view of the final-score advantage reported in the main paper.

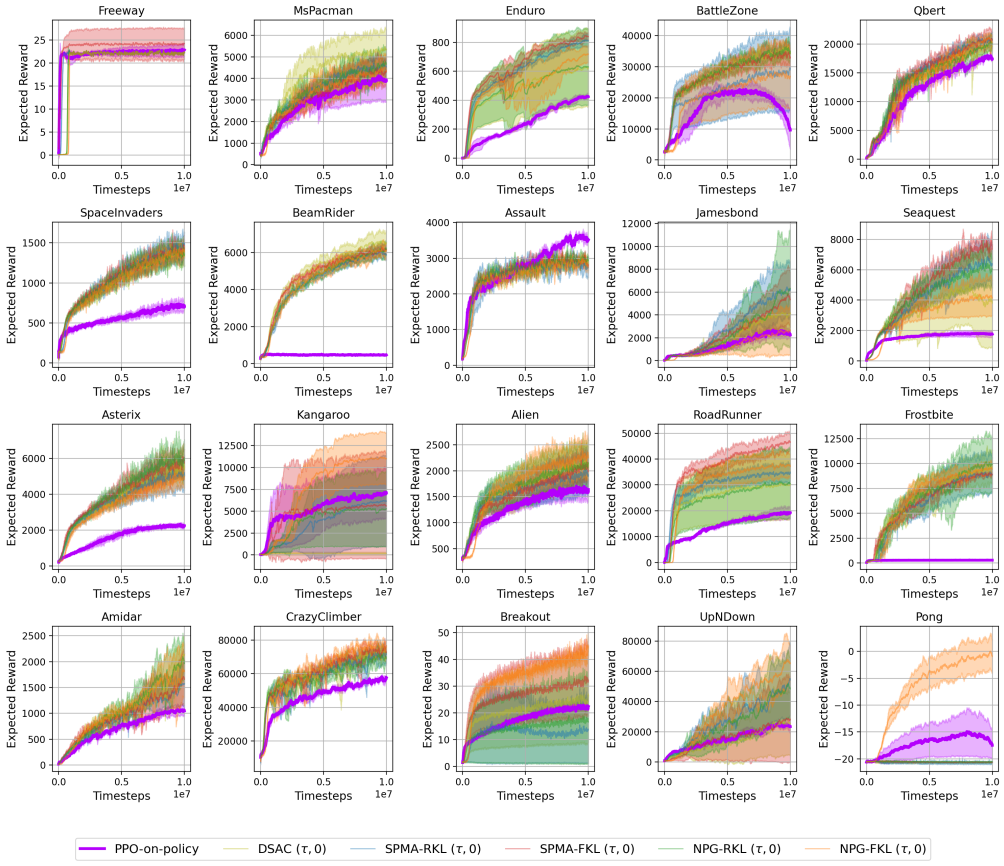


Figure B.14: On-policy `PPO` compared to objectives presented in this paper over the full training horizon. Overall, `PPO` underperforms.

Comparison with DQN. Figure B.15 compares the same objectives with DQN throughout training. Their advantage on most games shows that the main-paper gains are not limited to the final checkpoint and that the actor-critic objectives remain competitive with a strong value-based baseline over the full horizon.

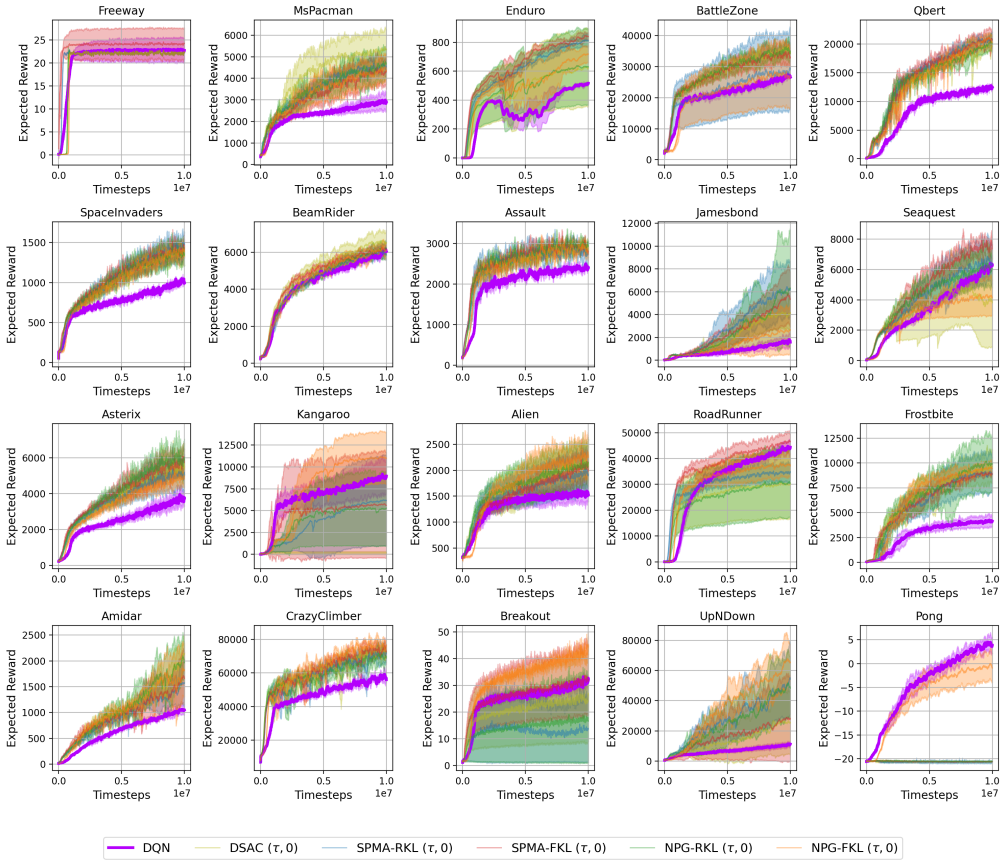


Figure B.15: DQN compared to objectives presented in this paper over the full training horizon. Overall, DQN underperforms on most games.

Comparison with Adam LMCDQN. Figure B.16 provides the full-horizon comparison with Adam LMCDQN, the strongest value-based baseline considered in the paper. The broadly comparable trajectories support the paper's conclusion that the proposed off-policy actor-critic framework can match a competitive value-based method while offering a unified family of policy objectives.

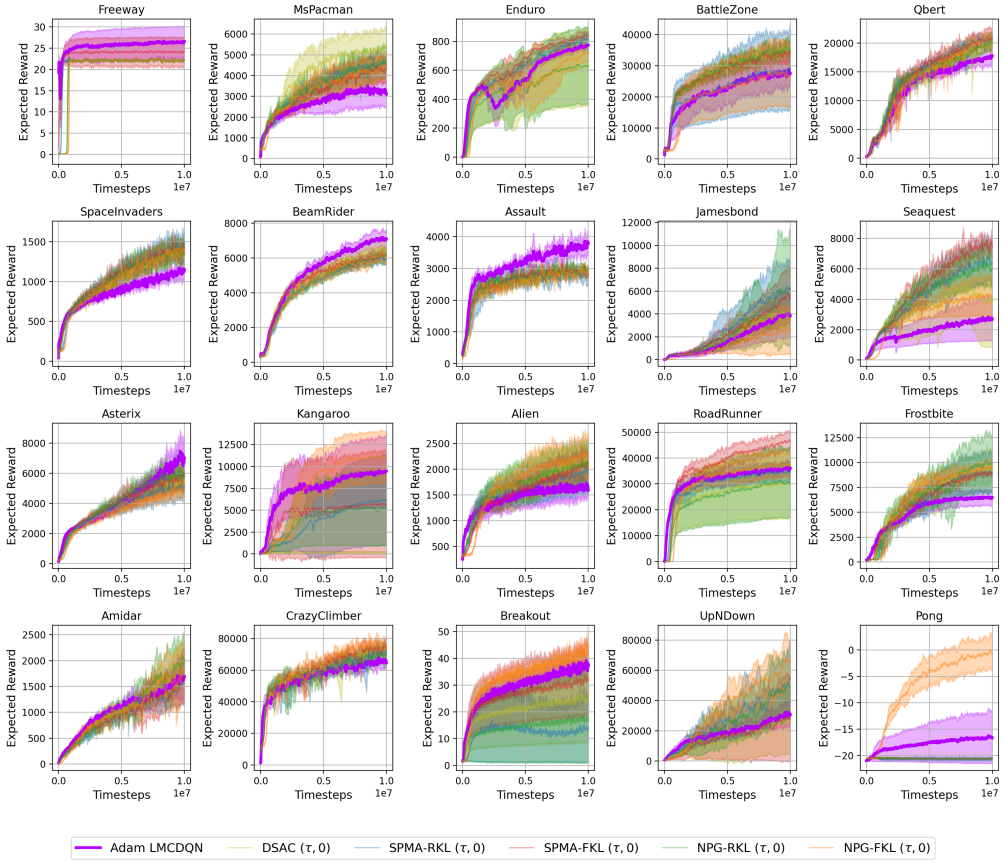


Figure B.16: Comparing `Adam LMCDQN` with all off-policy objectives presented in this paper over the full training horizon. Overall, the objectives from our proposed framework exhibit performance comparable to `Adam LMCDQN`.

B.5.11 Entropy Collapse of Default DSAC

When ablating the default DSAC on four Atari games in Section 3.5 (Figure 3.2), we find that it performs well on some games but fails on others (Alien and BeamRider). The results in Figure B.17 show that poor performance correlates with a collapse in policy entropy (columns 1 and 3, orange curve). Prior work [Xu et al., 2021] attributes this collapse to the fixed target entropy in DSAC, and our observations support this: reducing the learning rate for the entropy coefficient loss improves performance (blue curve). Nevertheless, removing critic entropy eliminates this issue, without requiring additional tuning of the entropy coefficient loss and results in strong performance.

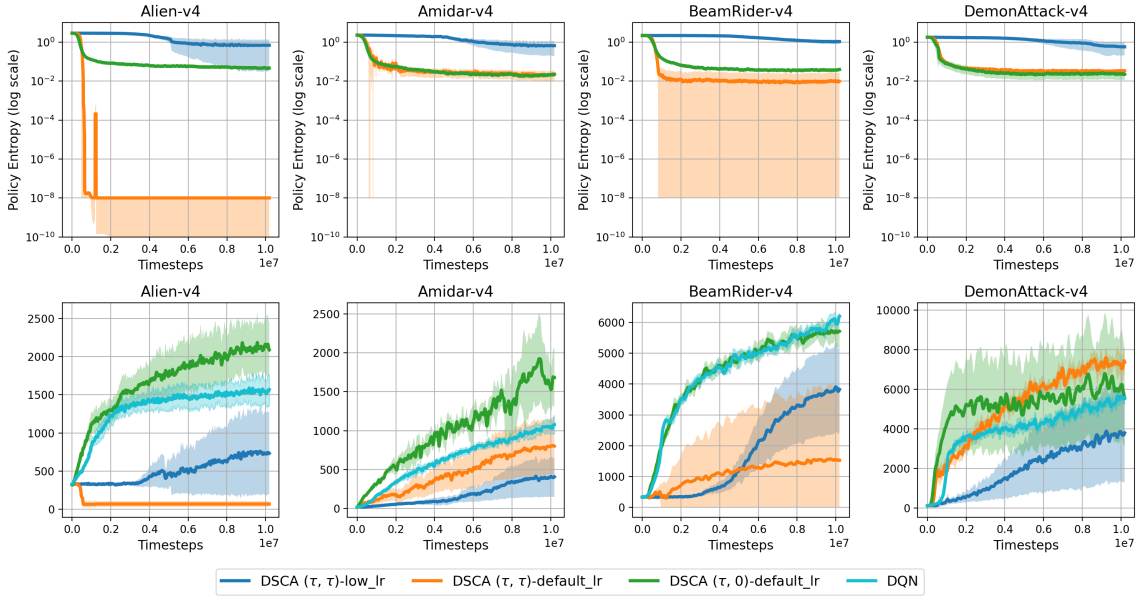


Figure B.17: The poor performance of default DSAC (orange curve) correlates with a collapse in policy entropy (columns 1 and 3). Although lowering the learning rate for the entropy coefficient loss improves performance, removing the critic entropy term resolves the issue without additional hyperparameter tuning.

Appendix C

Optimistic Actor-Critic with Parametric Policies

C.1 Notations

Notation	Meaning
Problem Definition	
$\mathcal{S}$, $\mathcal{A}$	state space and action space
H, h	horizon length (total number of steps), current index of step
$r \in \mathbb{R}^{\lvert\mathcal{S}\rvert\times\lvert\mathcal{A}\rvert}$	reward function
$\mathbb{P} \in \mathbb{R}^{\lvert\mathcal{S}\rvert\times\lvert\mathcal{A}\rvert\times\lvert\mathcal{S}\rvert}$	transition kernel
$\phi : \mathcal{S}\times\mathcal{A} \mapsto \mathbb{R}^{d_{\boldsymbol{c}}}$	features for the linear MDP environment
$\varphi : \mathcal{S}\times\mathcal{A} \mapsto \mathbb{R}^{d_{\boldsymbol{a}}}$	features for the learnable policy
Algorithm Design	
T, t	total number of learning episodes, index of current episode
$\mathcal{D}^t$, $\mathcal{D}_h^t$	collected data of at episode t, split data at h-th step (subset of $\mathcal{D}^t$)
N	number of samples collected for on-policy learning
$w \in \mathbb{R}^{d_{\boldsymbol{c}}}$	learnable critic parameters
J	number of critic updates
$\alpha_{\boldsymbol{c}}$	critic learning rate
ν	noise vector for LMC sampled from the standard normal distribution
ζ	inverse temperature for the `LMC` critic loss
M	number of samples for the critic parameters
$\mathcal{D}_{\text{exp}}$, ρ_{exp}	subset of $\mathcal{S}\times\mathcal{A}$, distribution over the subset
$\theta \in \mathbb{R}^{d_{\boldsymbol{a}}}$	learnable actor parameters
K	number of actor updates
$\alpha_{\boldsymbol{a}}$	actor learning rate
η	policy optimization learning rate

Table C.1: Notations for Problem Definition and Algorithm Design

Additional Notations. Throughout this paper, we use subscripts to represent the index of the step within the horizon of the episodic MDP and superscripts to denote the index of the episode for learning. For example, V_h^t means the value function for the h-th step derived at the learning episode t. In some cases, where the subscripts are omitted, it represents a set of H functions for all steps $h \in [H]$ (e.g., $V^t \coloneqq \{V_h^t\}_{h\in[H]}$). $|\mathcal{D}^t|$ represents the number of trajectories in $\mathcal{D}^t$ or the number of (s_h, a_h, s_{h+1}) tuples in $\mathcal{D}_h^t$. Additionally, for any vector $v \in \mathbb{R}^d$ and any matrix $M \in \mathbb{R}^{d\times d}$, we denote $\|v\|_M = \sqrt{v^\top M\, v}$.

C.2 Analysis for the Actor

C.2.1 Generalized OMD Regret (Proof of Theorem 6)

Proof of Theorem 6. Given the update of $p^{t+1/2}$ and the fact that $\Delta(\mathcal{A})$ is a convex set, we have the following optimality condition:

$$\left\langle u - p^{t+1/2}, -\eta\, g^t + \log(p^{t+1/2}) - \log(p^t) \right\rangle \geq 0\,. \tag{C.1}$$

Then, for each $t \in [T]$, we have that

$$\begin{aligned}
\left\langle u - p^t, \eta\, g^t \right\rangle &= \left\langle u - p^{t+1/2}, \eta\, g^t \right\rangle + \left\langle p^{t+1/2} - p^t, \eta\, g^t \right\rangle && \text{(C.2)}\\
&= \left\langle u - p^{t+1/2}, \eta\, g^t - \log(p^{t+1/2}) + \log(p^t) \right\rangle + \left\langle u - p^{t+1/2}, \log(p^{t+1/2}) - \log(p^t) \right\rangle && \text{(C.3)}\\
&\quad + \left\langle p^{t+1/2} - p^t, \eta\, g^t \right\rangle && \text{(C.4)}\\
&\overset{\text{(i)}}{\leq} \left\langle u - p^{t+1/2}, \log(p^{t+1/2}) - \log(p^t) \right\rangle + \left\langle p^{t+1/2} - p^t, \eta\, g^t \right\rangle && \text{(C.5)}\\
&\overset{\text{(ii)}}{=} \mathrm{KL}(u \,\|\, p^t) - \mathrm{KL}(u \,\|\, p^{t+1/2}) - \mathrm{KL}(p^{t+1/2} \,\|\, p^t) + \left\langle p^{t+1/2} - p^t, \eta\, g^t \right\rangle && \text{(C.6)}\\
&\overset{\text{(iii)}}{\leq} \mathrm{KL}(u \,\|\, p^t) - \mathrm{KL}(u \,\|\, p^{t+1/2}) - \mathrm{KL}(p^{t+1/2} \,\|\, p^t) + \frac{1}{2}\left\|p^{t+1/2} - p^t\right\|_1^2 + \frac{1}{2}\left\|\eta\, g^t\right\|_\infty^2 && \text{(C.7)}\\
&\overset{\text{(iv)}}{\leq} \mathrm{KL}(u \,\|\, p^t) - \mathrm{KL}(u \,\|\, p^{t+1/2}) - \mathrm{KL}(p^{t+1/2} \,\|\, p^t) + \mathrm{KL}(p^{t+1/2} \,\|\, p^t) + \frac{\eta^2\, H^2}{2} && \text{(C.8)}\\
&\leq \mathrm{KL}(u \,\|\, p^t) - \mathrm{KL}(u \,\|\, p^{t+1/2}) + \frac{\eta^2\, H^2}{2} && \text{(C.9)}\\
&\leq \mathrm{KL}(u \,\|\, p^t) - \mathrm{KL}(u \,\|\, p^{t+1}) + \mathrm{KL}(u \,\|\, p^{t+1}) - \mathrm{KL}(u \,\|\, p^{t+1/2}) + \frac{\eta^2\, H^2}{2} && \text{(C.10)}\\
&= \mathrm{KL}(u \,\|\, p^t) - \mathrm{KL}(u \,\|\, p^{t+1}) + \epsilon^t + \frac{\eta^2\, H^2}{2}\,. && \text{(C.11)}
\end{aligned}$$

(i) drops the first term due to the optimality condition from Equation (C.1). (ii) applies the three-point property of Bregman divergence (Lemma 12) by setting $x = u$, $y = p^{t+1/2}$, and $z = p^t$. (iii) follows from the Hölder's inequality and then the Young's inequality (i.e., $\langle u, v\rangle \leq \|u\|_1 \|v\|_\infty \leq \|u\|_1^2/2 + \|v\|_\infty^2/2$), and (iv) applies the Pinkster's inequality and

$|g^t| \leq H$. Summing up the above inequality from $t = 1$ to T yields that

$$\sum_{t=1}^{T} \left\langle u - p^t, \eta\, g^t \right\rangle = \sum_{t=1}^{T} \mathrm{KL}(u \parallel p^t) - \mathrm{KL}(u \parallel p^{t+1}) + \sum_{t=1}^{T} \epsilon^t + \frac{\eta^2\, H^2\, T}{2} \tag{C.12}$$

$$= \mathrm{KL}(u \parallel p^1) - \mathrm{KL}(u \parallel p^{T+1}) + \sum_{t=1}^{T} \epsilon^t + \frac{\eta^2\, H^2\, T}{2} \tag{C.13}$$

$$\overset{\text{(v)}}{\leq} \mathrm{KL}(u \parallel p^1) + \sum_{t=1}^{T} \epsilon^t + \frac{\eta^2\, H^2\, T}{2} \tag{C.14}$$

$$\leq \sum_{a \in \mathcal{A}} u(a) \log(u(a)) - \sum_{a \in \mathcal{A}} u(a)\, \log(p^1(a)) + \sum_{t=1}^{T} \epsilon^t + \frac{\eta^2\, H^2\, T}{2} \tag{C.15}$$

$$\overset{\text{(vi)}}{\leq} \log |\mathcal{A}| + \sum_{t=1}^{T} \epsilon^t + \frac{\eta^2\, H^2\, T}{2}\,. \tag{C.16}$$

(v) follows from the fact that KL-divergence is non-negative, and (vi) stands because the first term is negative, and for the second term, p^1 is a uniform distribution. Dividing both side by η, we have that

$$\sum_{t=1}^{T} \left\langle u - p^t, g^t \right\rangle \leq \frac{\log |\mathcal{A}| + \sum_{t=1}^{T} \epsilon^t}{\eta} + \frac{\eta\, H^2\, T}{2}\,. \tag{C.17}$$

This concludes the proof. □

C.2.2 Projection Error (Proof of Lemma 1)

Proof of Lemma 1. First, we define Φ as the log-sum-exp mirror map and $\Phi^\star$ as negative entropy, its Fenchel conjugate. Based on this, for any softmax policy π, we can also define its logit as $z \coloneqq \nabla\Phi^\star(\pi) = (\nabla\Phi)^{-1}(\pi)$. Consequently, $\pi = \nabla\Phi(z)$. Additionally, for any two softmax policies π, π' and their corresponding logits z, z', it holds that, $D_\Phi(z, z') = \mathrm{KL}(\pi', \pi)$. Since we are using the log-linear policy class, we have $z_h^t(s,a) = \left\langle \varphi(s,a), \widehat{\theta}_h^t \right\rangle$ for any $(s,a) \in \mathcal{S} \times \mathcal{A}$ where $\widehat{\theta}_h^t$ represents the parameters we attain at episode t. Therefore, for any $s \in \mathcal{S}$,

$$\epsilon_h^t(s) = \mathrm{KL}(\pi^\star(\cdot \mid s) \parallel \pi_h^{t+1}(\cdot \mid s)) - \mathrm{KL}(\pi_h^\star(\cdot \mid s) \parallel \pi_h^{t+1/2}(\cdot \mid s)) \tag{C.18}$$

$$= D_\Phi(z_h^{t+1}(\cdot \mid s), z_h^\star(s,\cdot)) - D_\Phi(z_h^{t+1/2}(\cdot \mid s), z_h^\star(\cdot \mid s)) \tag{C.19}$$

$$\overset{\text{(i)}}{=} \left\langle \nabla\Phi(z_h^{t+1}(s,\cdot))) - \nabla\Phi(z_h^\star(s,\cdot))), z_h^{t+1}(s,\cdot)) - z_h^{t+1/2}(\cdot \mid s)) \right\rangle - D_\Phi(z_h^{t+1/2}(\cdot \mid s), z_h^{t+1}(\cdot \mid s)) \tag{C.20}$$

$$= \left\langle \pi_h^{t+1}(s,\cdot) - \pi_h^\star(s,\cdot), z_h^{t+1}(s,\cdot)) - z_h^{t+1/2}(s,\cdot)) \right\rangle - \mathrm{KL}(\pi_h^{t+1}(\cdot \mid s)) \parallel \pi_h^{t+1/2}(\cdot \mid s))) \tag{C.21}$$

$$\overset{\text{(ii)}}{\leq} \left\langle \pi_h^{t+1}(\cdot \mid s) - \pi_h^\star(\cdot \mid s), z_h^{t+1}(s,\cdot)) - z_h^{t+1/2}(s,\cdot)) \right\rangle \tag{C.22}$$

$$\overset{\text{(iii)}}{\leq} \left\| \pi_h^{t+1}(\cdot \mid s) - \pi^\star(\cdot \mid s) \right\|_2 \left\| z_h^{t+1}(s,\cdot)) - z_h^{t+1/2}(s,\cdot)) \right\|_2 \tag{C.23}$$

$$\leq \sqrt{2}\, \left\| z_h^{t+1}(s,\cdot) - z_h^{t+1/2}(s,\cdot) \right\|_2 \tag{C.24}$$

$$\overset{\text{(iv)}}{=} \sqrt{2}\, \left\| \left\langle \varphi(s,\cdot), \widehat{\theta}_h^{t+1} - \widehat{\theta}_h^{t} \right\rangle - \eta\, \widehat{Q}_h^t(s,\cdot) \right\|_2 \tag{C.25}$$

$$\overset{\text{(v)}}{\leq} \sqrt{2}\, \left| \left\langle \varphi(s,\cdot), \widehat{\theta}_h^{t+1} - \widehat{\theta}_h^{t} \right\rangle - \eta\, \widehat{Q}_h^t(s,\cdot) \right| . \tag{C.26}$$

(i) follows from the three-point property of Bregman divergence (Lemma 12) by setting $x = z_h^{t+1/2}(\cdot \mid s)$, $y = z_h^{t+1}(\cdot \mid s)$, and $z = z_h^\star(\cdot \mid s)$ where $z^\star$ is the logit of $\pi^\star$. (ii) is based on the fact that KL-divergence is non-negative. (iii) uses the Cauchy-Schwarz inequality. (iv) uses the NPG update. (v) holds because $\|\cdot\|_2 \leq \|\cdot\|_1$.

Since the actor is designed to minimize the ridge regression in Algorithm 3, the minimizer can be written as

$$\widehat{\theta}_h^{t,\star} = \arg\min_{\theta_h} \frac{1}{2} \sum_{(s,a)\in\mathcal{D}_{\text{exp}}} \rho(s,a)\, \left[\langle \varphi(s,a), \theta_h \rangle - \widehat{Z}_h^t(s,a) \right]^2 , \tag{C.27}$$

where $\widehat{Z}_h^t(s,a) \coloneqq \left\langle \varphi(s,\cdot), \widehat{\theta}_h^t(s,\cdot) \right\rangle + \eta\, \widehat{Q}_h^t(s,a)$ for all $t \in [T]$. We define $\widehat{\theta}_h^{t,\star}$ as the minimizer, and it has the following explicit solution:

$$\widehat{\theta}_h^{t,\star} = G^{-1} \left[\sum_{(s',a')\in\mathcal{D}_{\text{exp}}} \rho(s',a')\, \widehat{Z}_h^t(s',a')\, \varphi(s',a') \right] , \tag{C.28}$$

where $G \coloneqq \sum_{(s,a)\in\mathcal{D}_{\text{exp}}} \rho(s,a)\, \varphi(s,a)\, \varphi(s,a)^\top \in \mathbb{R}^{d_{\boldsymbol{a}} \times d_{\boldsymbol{a}}}$.

Suppose $\theta_h^{t,\star}$ is the minimizer of the regression loss over the entire state-action space, $\widehat{\theta}_h^{t,\star}$ is the minimizer over the coreset, and $\widehat{\theta}_h^t$ is the parameters produced by the actor after K_t rounds of gradient descent as shown in Algorithm 3. Under Assumptions 4 and 5, we know that for any t and h,

$$\left| \left\langle \varphi(s,a), \widehat{\theta}_h^{t,\star} \right\rangle - \widehat{Z}_h^t(s,a)) \right| \leq \sqrt{2\, \epsilon_{\text{bias}}}\, , \tag{C.29}$$

$$\left| \left\langle \varphi(s,a), \widehat{\theta}_h^{t} \right\rangle - \left\langle \varphi(s,a), \widehat{\theta}_h^{t,\star} \right\rangle \right| \leq \sqrt{2\, \epsilon_{\text{opt}}}\, . \tag{C.30}$$

Then, for any arbitrary $(s,a) \in \mathcal{S} \times \mathcal{A}$, using the triangular inequality, we have that

$$\left| \left\langle \varphi(s,a), \widehat{\theta}_h^{t} \right\rangle - \widehat{Z}_h^t(s,a)) \right| \tag{C.31}$$

$$\leq \left| \left\langle \varphi(s,a), \theta_h^{t,\star} \right\rangle - \widehat{Z}_h^t(s,a)) \right| + \left| \left\langle \varphi(s,a), \widehat{\theta}_h^{t} \right\rangle - \left\langle \varphi(s,a), \theta_h^{t,\star} \right\rangle \right| \tag{C.32}$$

$$= \sqrt{2\, \epsilon_{\text{bias}}} + \left| \left\langle \varphi(s,a), \widehat{\theta}_h^{t} \right\rangle - \left\langle \varphi(s,a), \theta_h^{t,\star} \right\rangle \right| \tag{C.33}$$

$$\leq \sqrt{2\, \epsilon_{\text{bias}}} + \left| \left\langle \varphi(s,a), \widehat{\theta}_h^{t} \right\rangle - \left\langle \varphi(s,a), \widehat{\theta}_h^{t,\star} \right\rangle \right| + \left| \left\langle \varphi(s,a), \widehat{\theta}_h^{t,\star} \right\rangle - \left\langle \varphi(s,a), \theta_h^{t,\star} \right\rangle \right| \tag{C.34}$$

$$= \sqrt{2\, \epsilon_{\text{bias}}} + \sqrt{2\, \epsilon_{\text{opt}}} + \left| \left\langle \varphi(s,a), \widehat{\theta}_h^{t,\star} - \theta_h^{t,\star} \right\rangle \right| . \tag{C.35}$$

Therefore, it suffices to bound $\left|\left\langle \varphi(s,a), \widehat{\theta}_h^{t,\star}(s,a) - \theta_h^{t,\star}(s,a)\right\rangle\right|$. To do that, we first define $\Upsilon(s',a') \coloneqq \widehat{Z}_h^t(s',a') - \left\langle \varphi(s',a'), \theta_h^{t,\star}\right\rangle$ for any $(s',a') \in \mathcal{D}_{\text{exp}}$. Then, we have that

$$\widehat{\theta}_h^{t,\star} = G^{-1}\left[\sum_{(s',a')\in\mathcal{D}_{\text{exp}}} \rho(s',a')\,\left[\Upsilon(s',a') + \left\langle \varphi(s',a'), \theta_h^{t,\star}\right\rangle\right]\,\varphi(s',a')\right] \tag{C.36}$$

$$= G^{-1}\left[\sum_{(s',a')\in\mathcal{D}_{\text{exp}}} \rho(s',a')\,\varphi(s',a')\,\varphi(s',a')^\top\right]\theta_h^{t,\star} + G^{-1}\left[\sum_{(s',a')\in\mathcal{D}_{\text{exp}}} \rho(s',a')\,\Upsilon(s',a')\,\varphi(s',a')\right] \tag{C.37}$$

$$= \theta_h^{t,\star} + G^{-1}\left[\sum_{(s',a')\in\mathcal{D}_{\text{exp}}} \rho(s',a')\,\Upsilon(s',a')\,\varphi(s',a')\right]. \tag{C.38}$$

This implies that

$$\widehat{\theta}_h^{t,\star} - \theta_h^{t,\star} = G^{-1}\left[\sum_{(s',a')\in\mathcal{D}_{\text{exp}}} \rho(s',a')\,\Upsilon(s',a')\,\varphi(s',a')\right]. \tag{C.39}$$

Hence, for any arbitrary $(s,a) \in \mathcal{S}\times\mathcal{A}$,

$$\left|\left\langle \varphi(s,a), \widehat{\theta}_h^{t,\star} - \theta_h^{t,\star}\right\rangle\right| = \left|\sum_{(s',a')\in\mathcal{D}_{\text{exp}}} \rho(s',a')\,\Upsilon(s',a')\varphi(s,a)^\top\,G^{-1}\,\varphi(s',a')\right| \tag{C.40}$$

$$\overset{\text{(vi)}}{\leq} \sum_{(s',a')\in\mathcal{D}_{\text{exp}}} \left|\Upsilon(s',a')\right|\,\rho(s',a')\,\left|\varphi(s,a)^\top\,G^{-1}\,\varphi(s',a')\right| \tag{C.41}$$

$$\leq \left(\max_{(s',a')\in\mathcal{D}_{\text{exp}}} \left|\Upsilon(s',a')\right|\right)\sum_{(s',a')\in\mathcal{D}_{\text{exp}}} \rho(s',a')\,\left|\varphi(s,a)^\top\,G^{-1}\,\varphi(s',a')\right| \tag{C.42}$$

$$\leq \sqrt{2\,\epsilon_{\text{bias}}}\sum_{(s',a')\in\mathcal{D}_{\text{exp}}} \rho(s',a')\,\left|\varphi(s,a)^\top\,G^{-1}\,\varphi(s',a')\right| \tag{C.43}$$

$$= \sqrt{2\,\epsilon_{\text{bias}}}\sqrt{\left(\mathbb{E}_{(s',a')\sim\rho}\,\left|\varphi(s,a)^\top\,G^{-1}\,\varphi(s',a')\right|\right)^2} \tag{C.44}$$

$$\overset{\text{(vii)}}{\leq} \sqrt{2\,\epsilon_{\text{bias}}}\sqrt{\mathbb{E}_{(s',a')\sim\rho}\,\left|\varphi(s,a)^\top\,G^{-1}\,\varphi(s',a')\right|^2} \tag{C.45}$$

$$= \sqrt{2\,\epsilon_{\text{bias}}}\sqrt{\varphi(s,a)^\top G^{-1}\left[\sum_{(s',a')\in\mathcal{D}_{\text{exp}}} \rho(s',a')\,\varphi(s',a')\,\varphi(s',a')^\top\right]G^{-1}\varphi(s,a)} \tag{C.46}$$

$$= \sqrt{2\,\epsilon_{\text{bias}}}\,\left\|\varphi(s,a)\right\|_{G^{-1}}. \tag{C.47}$$

(vi) applies the Cauchy-Schwarz inequality, and (vii) follows from Jensen's inequality. Putting everything together, we have that

$$\left|\left\langle \varphi(s,a), \widehat{\theta}_h^t\right\rangle - \widehat{Z}_h^t(s,a))\right| \leq \sqrt{2}\left[\left(\left\|\varphi(s,a)\right\|_{G^{-1}} + 1\right)\sqrt{\epsilon_{\text{bias}}} + \sqrt{\epsilon_{\text{opt}}}\right] \tag{C.48}$$

$$\leq \sqrt{2}\left[\left(\overline{\varphi}_G + 1\right)\sqrt{\epsilon_{\text{bias}}} + \sqrt{\epsilon_{\text{opt}}}\right]. \tag{C.49}$$

Recall that $\epsilon_h^t(s) \le \sqrt{2}\,\left\|\left\langle \varphi(s,\cdot), \widehat{\theta}_h^{t+1}(s,\cdot)\right\rangle - \widehat{Z}_h^t(s,\cdot)\right\|_2$. Therefore, for any $s \in \mathcal{S}$,

$$\epsilon_h^t(s) \le \sqrt{2}\,\left|\left\langle \varphi(s,\cdot), \widehat{\theta}_h^t\right\rangle - \widehat{Z}_h^t(s,a))\right| \le 2\,(\overline{\varphi}_G + 1)\,\sqrt{\epsilon_{\text{bias}}} + 2\,\sqrt{\epsilon_{\text{opt}}}\,. \tag{C.50}$$

This concludes the proof. □

C.2.3 Instantiating the Actor with SPMA

Lemma 1 can not only be applied to NPG but also other mirror descent-based policy optimization methods such as MDPO [Tomar et al., 2020] and SPMA [Asad et al., 2025b]. In this section, as an example, we show that the projected variant of SPMA (projected SPMA) is also compatible with our framework and can enjoy similar sample complexity guarantees as projected NPG.

We can instantiate the actor in Algorithm 2 with the projected SPMA by setting the actor loss in Algorithm 3 as

$$\tilde{\ell}_h^t(\theta) = \frac{1}{2} \sum_{(s,a)\in\mathcal{D}_{\text{exp}}} \rho_{\text{exp}}(s,a)\,\left[\langle \varphi(s,a),\theta\rangle - \widehat{Z}_h^t(s,a)\right]^2, \tag{C.51}$$

$$\text{where } \widehat{Z}_h^t(s,a) \coloneqq \left\langle \varphi(s,a), \widehat{\theta}_h^t\right\rangle + \log(1+\eta\,A^{\pi^t}(s,\cdot))\,. \tag{C.52}$$

Equivalently, the projected SPMA update can be expressed as follows. For any $s \in \mathcal{S}$, $\pi^1(\cdot \mid s)$ is a uniform distribution, and

$$\pi^{t+1/2}(\cdot \mid s) = \arg\min_{p\in\Delta_{\mathcal{A}}} \left\{\left\langle \pi^t(\cdot \mid s), -\log\left(1+\eta\,A^{\pi^t}(s,\cdot)\right)\right\rangle + \text{KL}\left(p \parallel \pi^t(\cdot \mid s)\right)\right\}, \tag{C.53}$$

$$\pi^{t+1}(\cdot \mid s) = \text{Proj}_{\Pi}(\pi^{t+1/2}(\cdot \mid s))\,. \tag{C.54}$$

Hence, we introduce the following alternative lemma to show that Theorem 6 also holds for the projected SPMA.

Lemma 11. *Given a sequence of linear functions $\{\langle p^t, g^t\rangle\}_{t\in[T]}$ for a sequence of vectors $\{g^t\}_{t\in[T]}$ where for any $t \in [T]$, $p^t \in \Delta(\mathcal{A})$, $g^t \in \mathbb{R}^{|\mathcal{A}|}$, and $g^t(a) \in [0,H]$ for all $a \in \mathcal{A}$. Consider $p^{t\in[T]}$ where p^1 is the uniform distribution, and for all $t \in [T]$,*

$$p^{t+1/2} = \arg\min_{p\in\Delta_A} \left\{\left\langle p, -\log\left(1+\eta\,\left(g^t - \left\langle p^t, g^t\right\rangle\,\mathbf{1}\right)\right)\right\rangle + \text{KL}(p \parallel p^t)\right\}, \tag{C.55}$$

$$p^{t+1} = \text{Proj}_{\Pi}(p^{t+1/2})\,, \tag{C.56}$$

where $\mathbf{1} \in \mathbb{R}^{|\mathcal{A}|}$ is an all-one vector. Let $\epsilon^t \coloneqq \text{KL}(u \parallel p^{t+1}) - \text{KL}(u \parallel p^{t+1/2})$ be the projection error induced by Equation (4.7). If $\eta \le \frac{1}{2\,H}$, then for any comparator $u \in \Delta(\mathcal{A})$, it holds that

$$\sum_{t=1}^{T} \left\langle u - p^t, g^t\right\rangle \le \frac{\log|\mathcal{A}| + \sum_{t=1}^T \epsilon^t}{\eta} + \frac{3\,\eta\,H^2\,T}{2}\,. \tag{C.57}$$

Proof of Lemma 11. We first denote that $d^t = \log\left(1+\eta\,\left(g^t - \langle p^t, g^t\rangle\,\mathbf{1}\right)\right)$ for all $t \in [T]$. Then, for all $a \in \mathcal{A}$, since $\eta \le \frac{1}{2\,H}$ and $g^t(a) - \langle p^t, g^t\rangle \in [-H, H]$, we have $\eta\,\left(g^t(a) - \langle p^t, g^t\rangle\right) >$

$-\frac{1}{2}$ and therefore

$$d^t(a) \overset{\text{(i)}}{\leq} \eta \left(g^t(a) - \left\langle p^t, g^t \right\rangle \right) \leq \eta H \,, \tag{C.58}$$

$$d^t(a) \overset{\text{(ii)}}{\geq} \eta \left(g^t(a) - \left\langle p^t, g^t \right\rangle \right) - \eta^2 \left(g^t(a) - \left\langle p^t, g^t \right\rangle \right)^2 \geq \eta \left(g^t(a) - \left\langle p^t, g^t \right\rangle \right) - \eta H^2 \,, \tag{C.59}$$

where (i) follows from $\log(1+x) \leq x$ for all $x > -1$, and (ii) holds because $\log(1+x) \geq x - x^2$ for all $x > -\frac{1}{2}$.

Given the update of $p^{t+1/2}$ and the fact that $\Delta(\mathcal{A})$ is a convex set, we have the following optimality condition:

$$\left\langle u - p^{t+1/2}, -d^t + \log(p^{t+1/2}) - \log(p^t) \right\rangle \geq 0 \,. \tag{C.60}$$

Then, for all $t \in [T]$, we have that

$$\left\langle u - p^t, d^t \right\rangle = \left\langle u - p^{t+1/2}, d^t \right\rangle + \left\langle p^{t+1/2} - p^t, d^t \right\rangle \tag{C.61}$$

$$= \left\langle u - p^{t+1/2}, d^t - \log(p^{t+1/2}) + \log(p^t) \right\rangle + \left\langle u - p^{t+1/2}, \log(p^{t+1/2}) - \log(p^t) \right\rangle \tag{C.62}$$

$$+ \left\langle p^{t+1/2} - p^t, d^t \right\rangle \tag{C.63}$$

$$\overset{\text{(iii)}}{\leq} \left\langle u - p^{t+1/2}, \log(p^{t+1/2}) - \log(p^t) \right\rangle + \left\langle p^{t+1/2} - p^t, d^t \right\rangle \tag{C.64}$$

$$\overset{\text{(iv)}}{=} \mathrm{KL}(u \,\|\, p^t) - \mathrm{KL}(u \,\|\, p^{t+1/2}) - \mathrm{KL}(p^{t+1/2} \,\|\, p^t) + \left\langle p^{t+1/2} - p^t, d^t \right\rangle \tag{C.65}$$

$$\overset{\text{(v)}}{\leq} \mathrm{KL}(u \,\|\, p^t) - \mathrm{KL}(u \,\|\, p^{t+1/2}) - \mathrm{KL}(p^{t+1/2} \,\|\, p^t) + \frac{1}{2} \left\| p^{t+1/2} - p^t \right\|_1^2 + \frac{1}{2} \left\| d^t \right\|_\infty^2 \tag{C.66}$$

$$\overset{\text{(vi)}}{\leq} \mathrm{KL}(u \,\|\, p^t) - \mathrm{KL}(u \,\|\, p^{t+1/2}) - \mathrm{KL}(p^{t+1/2} \,\|\, p^t) + \mathrm{KL}(p^{t+1/2} \,\|\, p^t) + \frac{\eta^2 H^2}{2} \tag{C.67}$$

$$\leq \mathrm{KL}(u \,\|\, p^t) - \mathrm{KL}(u \,\|\, p^{t+1/2}) + \frac{\eta^2 H^2}{2} \tag{C.68}$$

$$\leq \mathrm{KL}(u \,\|\, p^t) - \mathrm{KL}(u \,\|\, p^{t+1}) + \mathrm{KL}(u \,\|\, \pi^{t+1}) - \mathrm{KL}(u \,\|\, p^{t+1/2}) + \frac{\eta^2 H^2}{2} \tag{C.69}$$

$$= \mathrm{KL}(u \,\|\, p^t) - \mathrm{KL}(u \,\|\, p^{t+1}) + \epsilon^t + \frac{\eta^2 H^2}{2} \,. \tag{C.70}$$

(iii) drops the first term due to the optimality condition from Equation (C.60). (iv) applies the three-point property of Bregman divergence (Lemma 12) by setting $x = u$, $y = p^{t+1/2}$, and $z = p^t$. (v) follows from the Hölder's inequality and then the Young's inequality (i.e., $\langle u, v \rangle \leq \|u\|_1 \|v\|_\infty \leq \|u\|_1^2 / 2 + \|v\|_\infty^2 / 2$), and (vi) applies the Pinkster's inequality and $\|d^t\|_\infty \leq H$.

Moreover, we have that

$$\left\langle u - p^t, d^t \right\rangle \overset{\text{(vii)}}{\geq} \left\langle u - p^t, \eta \left(g^t - \left\langle p^t, g^t \right\rangle \mathbf{1} \right) \right\rangle - \eta H^2 \overset{\text{(viii)}}{\geq} \left\langle u - p^t, \eta \, g^t \right\rangle - \eta H^2 \,, \tag{C.71}$$

where (vii) comes from the fact that $d^t(a) \geq \eta\ (g^t(a) - \langle p^t, g^t \rangle) - \eta\, H^2$ for all $a \in \mathcal{A}$, and (viii) follows from the fact that $\langle p^t, g^t \rangle \geq 0$ since $p^t \in \Delta(\mathcal{A})$ and $g^t(a) \in [0, H]$ for all $a \in \mathcal{A}$. This implies that

$$\left\langle u - p^t, \eta\, g^t \right\rangle \leq \left\langle u - p^t, d^t \right\rangle + \eta\, H^2 \tag{C.72}$$

$$\leq \mathrm{KL}(u \parallel p^t) - \mathrm{KL}(u \parallel p^{t+1}) + \epsilon^t + \frac{3\,\eta^2\, H^2}{2}\,. \tag{C.73}$$

Summing up the above inequality from $t = 1$ to T yields that

$$\sum_{t=1}^{T} \left\langle u - p^t, \eta\, g^t \right\rangle = \sum_{t=1}^{T} \mathrm{KL}(u \parallel p^t) - \mathrm{KL}(u \parallel p^{t+1}) + \sum_{t=1}^{T} \epsilon^t + \frac{3\,\eta^2\, H^2\, T}{2} \tag{C.74}$$

$$= \mathrm{KL}(u \parallel p^1) - \mathrm{KL}(u \parallel p^{T+1}) + \sum_{t=1}^{T} \epsilon^t + \frac{3\,\eta^2\, H^2\, T}{2} \tag{C.75}$$

$$\overset{\text{(ix)}}{\leq} \mathrm{KL}(u \parallel p^1) + \sum_{t=1}^{T} \epsilon^t + \frac{3\,\eta^2\, H^2\, T}{2} \tag{C.76}$$

$$\leq \sum_{a \in \mathcal{A}} u(a) \log(u(a)) - \sum_{a \in \mathcal{A}} u(a)\, \log(p^1(a)) + \sum_{t=1}^{T} \epsilon^t + \frac{3\,\eta^2\, H^2\, T}{2} \tag{C.77}$$

$$\overset{\text{(x)}}{\leq} \log |\mathcal{A}| + \sum_{t=1}^{T} \epsilon^t + \frac{3\,\eta^2\, H^2\, T}{2}\,. \tag{C.78}$$

(ix) follows from the fact that KL-divergence is non-negative, and (x) stands because the first term is negative, and for the second term, p^1 is a uniform distribution. Dividing both side by η, we have that

$$\sum_{t=1}^{T} \left\langle u - p^t, g^t \right\rangle \leq \frac{\log |\mathcal{A}| + \sum_{t=1}^{T} \epsilon^t}{\eta} + \frac{3\,\eta\, H^2\, T}{2}\,. \tag{C.79}$$

This concludes the proof. □

In order to obtain a meaningful regret bound, we should set $\eta = \min\left\{\frac{1}{2\,H}, \sqrt{\frac{2\,(\log |\mathcal{A}| + \bar{\epsilon}\, T)}{3\, H^2\, T}}\right\}$. Therefore, under Assumptions 4 and 5, we can easily prove that Lemma 1 also holds for the projected SPMA, and consequently, all the sample complexity guarantees for the projected NPG should also hold.

C.2.4 Technical Tools

Lemma 12 (Three-Point Property of Bregman Divergence)**.** *Suppose $X \subseteq \mathbb{R}^d$ is closed and convex. Consider a strictly convex function $\Phi : X \to \mathbb{R}$. For all $x \in X$ and $y, z \in intX$,*

$$D_\Phi(x, y) + D_\Phi(y, z) - D_\Phi(x, z) = \langle \nabla\Phi(z) - \nabla\Phi(y), x - y \rangle\ . \tag{C.80}$$

C.3 Constructing $\mathcal{D}_{\text{exp}}$ and ρ_{exp} via Experimental Design

In this section, we introduce various methods of experimental design to bound $\overline{\varphi}_G$ defined in Lemma 1. The experimental design problem can be written as

$$\inf_{\substack{\mathcal{D}_{\text{exp}} \in \mathcal{S}\times\mathcal{A} \\ \rho_{\text{exp}} \in \Delta(\mathcal{D}_{\text{exp}})}} \sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} \|\varphi(s,a)\|_{G^{-1}} \tag{C.81}$$

$$\text{s.t.} \quad G = \sum_{(s,a)\in\mathcal{D}_{\text{exp}}} \rho_{\text{exp}}(s,a)\, \varphi(s,a)\, \varphi(s,a)^\top . \tag{C.82}$$

In Section C.3.1, we consider constructing a coreset for the policy features. The Kiefer–Wolfowitz theorem guarantees that there exists a coreset that can ensure that $\overline{\varphi}_G$ is bounded, and that such a coreset has a small $O(d)$ size. Such a coreset can be formed using G-experimental design. In Section C.3.2, we consider using the linear MDP features as the policy features and constructing $\mathcal{D}_{\text{exp}}$ through limited interaction with the environment.

C.3.1 Kiefer–Wolfowitz Theorem and G-Experimental Design

We first introduce the Kiefer–Wolfowitz theorem [Kiefer and Wolfowitz, 1960] which guarantees that there exists a coreset $\mathcal{D}_{\text{exp}}$ and its corresponding distribution ρ_{exp} that can be used to bound $\overline{\varphi}_G$.

Proposition 2 (Kiefer–Wolfowitz)**.** *Let $G \coloneqq \sum_{(s,a)\in\mathcal{D}_{\text{exp}}} \rho_{\text{exp}}(s,a)\, \varphi(s,a)\, \varphi(s,a)^\top$ be the covariance matrix for any $\mathcal{D}_{\text{exp}} \subset \mathcal{S}\times\mathcal{A}$ and $\rho \in \Delta(\mathcal{D}_{\text{exp}})$. There exists a coreset $\mathcal{D}_{\text{exp}}$ and a distribution ρ_{exp} such that*

$$\sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} \|\varphi(s,a)\|_{G^{-1}} \le 2\, d_{\boldsymbol{a}} \quad \textit{and} \quad |\mathcal{D}_{\text{exp}}| \le 4 d_{\boldsymbol{a}} \log\log(d_{\boldsymbol{a}} + 4) + 28 . \tag{C.83}$$

Note that the size of $\mathcal{D}_{\text{exp}}$ is also bounded by $\widetilde{\mathcal{O}}(d_{\boldsymbol{a}})$, suggesting that the computation cost of calculating the actor loss over $\mathcal{D}_{\text{exp}}$ is inexpensive. The problem of constructing such a coreset is often framed as G-experimental design, and it can typically be solved using numerous efficient approximation algorithms such as the Franke-Wolfe algorithm [Frank et al., 1956] as mentioned in Todd [2016], Lattimore and Szepesvári [2020]. Using $\mathcal{D}_{\text{exp}}$ and ρ_{exp} produced by such methods to construct the actor loss in Algorithm 3 offers the guarantees that $\overline{\varphi}_G \le \mathcal{O}(d_{\boldsymbol{a}})$, which is consequently used to bound the projection error in Lemma 1 as $\bar{\epsilon} \le \mathcal{O}(d_{\boldsymbol{a}}\, \epsilon_{\text{bias}} + \epsilon_{\text{opt}})$.

We remark that the coreset construction can be done before the learning process in the actor-critic algorithm since it is independent of the linear MDP environment. However, these algorithms typically require traversing through all the policy features in $\mathcal{S}\times\mathcal{A}$, which is not ideal for large state-action spaces.

C.3.2 Exploratory Policy and Minimum Eigenvalue

Alternatively, we can choose to use the linear MDP features as the policy features (i.e., $\varphi = \phi$) and construct $\mathcal{D}_{\text{exp}}$ via interacting with the environment. Note that bounding $\overline{\varphi}_G$ is equivalent to controlling $\|\phi(s,a)\|_{G^{-1}}$ for all $(s,a) \in \mathcal{S}\times\mathcal{A}$. Consequently, given that

$\|\phi(s,a)\|_2 \leq 1$ by the linear MDP assumption and since

$$\|\phi(s,a)\|_{G^{-1}} \leq \frac{\|\phi(s,a)\|_2}{\lambda_{\min}(G)} = \frac{1}{\lambda_{\min}(G)}\,, \tag{C.84}$$

we only need a well-conditioned covariance matrix G that has a positive minimum eigenvalue.

Several existing works [Hao et al., 2021, Agarwal et al., 2021b] assume access to an exploratory (not necessarily optimal) policy π_{exp} that is able to collect such covariance matrices with minimum eigenvalue bounded away from 0. Given that, we can directly apply π_{exp} to roll-out trajectories and collect observations, which can be used to construct $\mathcal{D}_{\text{exp}}$ and the corresponding covariance G.

However, in practice, we rarely have access to such an oracle policy. Consequently, Wagenmaker et al. [2022] proposed a reward-free approach, `CoverTraj`, that can effectively collect such observations without assuming access to an exploratory policy. In particular, the `CoverTraj` algorithm offers the following theoretical guarantee.

Proposition 3 (Wagenmaker et al. 2022, Theorem 4)**.** *Fix $h \in [H]$ and $\gamma \in [0,1]$. Suppose there exists a problem-dependent constant $\epsilon_{\mathcal{M}} > 0$ such that $\sup_{\pi\in\Pi} \lambda_{\min}\left(\mathbb{E}_\pi\left[\phi(s,a)\,\phi(s,a)^\top\right]\right) \geq \epsilon_{\mathcal{M}}$. Running K rounds of* `CoverTraj` *to collect $\mathcal{D}_{\text{exp}} = \{(s_h^\tau, a_h^\tau)\}_{\tau=1}^K$ where*

$$K = \widetilde{\mathcal{O}}\left(\frac{1}{\epsilon_{\mathcal{M}}} \cdot \max\left\{\frac{d_{\boldsymbol{c}}}{\gamma^2}, d_{\boldsymbol{c}}^4\, H^3\, \log^3\frac{1}{\delta}\right\}\right) \tag{C.85}$$

ensures that for any $\delta \in (0,1)$, with probability of at least $1-\delta$,

$$\lambda_{\min}(G) \geq \frac{\epsilon_{\mathcal{M}}}{\gamma^2}\,, \tag{C.86}$$

where $G = \sum_{(s,a)\in\mathcal{D}_{\text{exp}}} \phi(s,a)\,\phi(s,a)^\top$.

Note that `CoverTraj` does not utilize the reward function of the MDP and merely use the transition kernel when interacting with the environment. Alternatively, Wagenmaker and Jamieson [2022] provides another approach, `OptCov`, that utilizes regret minimization algorithms to construct the desired covariance matrix. According to Wagenmaker and Jamieson [2022, Theorem 9], `OptCov` can also offer a similar guarantee of the minimum eigenvalue ensuring that

$$\lambda_{\min}(G) \geq \max\left\{d_{\boldsymbol{c}}\,\log\left(\frac{1}{\delta}\right), \epsilon_{\mathcal{M}}\right\}. \tag{C.87}$$

To conclude, the Frank-Wolfe algorithm can be used to form a coreset and subsequently bound $\overline{\varphi}_G$ for any given policy features. If we use the linear MDP features as the policy features, we can construct $\mathcal{D}_{\text{exp}}$ by interacting with the environment. Either having access to an exploratory policy or running `CoverTraj` or `OptCOv` can offer guarantees on the minimum eigenvalues of the covariance matrix, which will consequently control $\overline{\varphi}_G$.

C.4 Analysis for the Critic

C.4.1 Proof of Lemma 2

In order to prove Lemma 2, we introduce the following "good" event for the estimated value function.

Lemma 13 (Good Event)**.** *There exists some $C_\delta > 0$ such that for any fixed $\delta \in (0,1)$, the following event,*

$$\mathcal{E}_\delta \coloneqq \left\{ \forall (t,h) \in [T] \times [H] : \left\| \sum_{(s,a,s') \in \mathcal{D}_h^t} \phi(s,a) \left[\widehat{V}_{h+1}^t(s') - \mathbb{P}_h \widehat{V}_{h+1}^t(s,a) \right] \right\|_{(\Lambda_h^t)^{-1}} \leq C_\delta \, H \, \sqrt{d_{\boldsymbol{c}}} \right\}, \tag{C.88}$$

holds with probability at least $1-\delta$ (i.e., $\Pr(\mathcal{E}_\delta) \geq 1-\delta$).

The exact definition of C_δ varies between the on-policy and the off-policy settings. We will prove that $\Pr(\mathcal{E}_\delta) \geq 1-\delta$ for the on-policy and off-policy setting in Section C.5 and Section C.6 respectively.

Next, conditioned on the above event, we present a formal version of Lemma 2, which provides an upper and a lower bound for the model prediction error induced by the LMC critic.

Lemma 14 (Formal version of Lemma 2)**.** *Consider Algorithm 2 with the* LMC *critic from Algorithm 4. Conditioned on $\mathcal{E}_\delta$ defined in Lemma 13, if we choose that $\lambda = 1$, $\zeta = \left(2\,H\,\sqrt{d_{\boldsymbol{c}}}\,C_\delta + 8/3\right)^{-2}$, $\alpha_{\boldsymbol{c}}^{h,t} = 1/\left(2\,\lambda_{\max}(\Lambda_h^t)\right.$, $J_t \geq 2\,\kappa_t\,\log\left(1/\sigma\right)$, and $M = \log\left(H\,T/\delta\right)/\log\left(1/(1-c)\right)$ where $\kappa_t = \max_{h\in[H]} \lambda_{\max}(\Lambda_h^t)/\lambda_{\min}(\Lambda_h^t)$, $\sigma = 1/\left(4\,H\,\left(\left|\mathcal{D}^t\right|+1\right)\,\sqrt{d_{\boldsymbol{c}}}\right)$, and $c = 1/(2\sqrt{2e\pi})$, then, for all $(t,h,s,a) \in [T]\times[H]\times\mathcal{S}\times\mathcal{A}$ and for any $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$-\Gamma_{\texttt{LMC}} \times \|\phi(s,a)\|_{(\Lambda_h^t)^{-1}} \leq \iota_h^t(s,a) \leq 0\,, \tag{C.89}$$

where $\Gamma_{\texttt{LMC}} = C_\delta\,H\,\sqrt{d_{\boldsymbol{c}}} + \frac{4}{3}\sqrt{\frac{2\,d_{\boldsymbol{c}}\log\left(1/\delta\right)}{3\,\zeta}} + \frac{4}{3} \leq \mathcal{O}\left(C_\delta\,H\,d_{\boldsymbol{c}}\sqrt{\log(1/\delta)}\right)$.

Preliminary Properties

In this section, we introduce some useful properties of LMC and state the supporting lemmas that will be helpful in proving the above result.

First, we obtain the derivative of the critic loss defined in Algorithm 4.

$$\nabla L_h^t(w_h) = \Lambda_h^t w_h - b_h^t, \tag{C.90}$$

where $\Lambda_h^t \coloneqq \sum_{(s,a)\in\mathcal{D}_h^t} \phi(s,a)\,\phi(s,a)^\top + \lambda\,I$ and $b_h^t \coloneqq \sum_{(s,a,s')\in\mathcal{D}_h^t} \left[r_h(s,a) + \widehat{V}_{h+1}^t(s')\right]\phi(s,a)$. Consequently, by setting $\nabla L_h^t(w_h) = 0$, we get the minimizer of $L_h^t(w_h)$ as

$$\widehat{w}_h^t \coloneqq (\Lambda_h^t)^{-1}\,b_h^t\,. \tag{C.91}$$

We now introduce the following lemma, showing that the noisy gradient descent performed by the LMC critic ensures that the sampled critic parameter w follows a Gaussian distribution.

Lemma 15 (Ishfaq et al. 2024a, Proposition B.1)**.** *Consider Algorithm 2 with the* LMC *critic from Algorithm 4. For any $(t,h,m) \in [T]\times[H]\times[M]$, the sampled parameters w_h^{t,m,J_t} follows a Gaussian distribution $\mathbf{N}\left(\mu_h^{t,m,J_t}, \Sigma_h^{t,m,J_t}\right)$. The mean and the covariance are defined as*

$$\mu_h^{t,J_t} = A_t^{J_t} \dots A_1^{J_1}\, w^{1,0} + \sum_{i=1}^{t} A_t^{J_t} \dots A_{i+1}^{J_{i+1}} (I - A_i^{J_i})\, \widehat{w}_h^i\,, \tag{C.92}$$

$$\Sigma_h^{t,J_t} = \frac{1}{\zeta} \sum_{i=1}^{t} A_t^{J_t} \dots A_{i+1}^{J_{i+1}}\, (I - A_i^{2\,J_i})\, (\Lambda_h^i)^{-1}\, (I + A_i)^{-1}\, A_{i+1}^{J_{i+1}} \dots A_t^{J_t}\,, \tag{C.93}$$

where $A_t \coloneqq I - \alpha_{\boldsymbol{c}}^t\, \Lambda_h^t$ for all $t \in [T]$.

Since w_h^{t,m,J_t} follows the Gaussian distribution of $\mathbf{N}\left(\mu_h^{t,m,J_t}, \Sigma_h^{t,m,J_t}\right)$, $\left\langle \phi_h(s,a), w_h^{t,m,J_t} \right\rangle$ also follows the Gaussian distribution of $\mathbf{N}\left(\phi_h(s,a)^\top\, \mu_h^{t,m,J_t}, \phi_h(s,a)^\top\, \Sigma_h^{t,m,J_t}\, \phi_h(s,a)\right)$. Therefore, we introduce the following lemmas to bound the terms related to the mean and variance.

Lemma 16. *Consider Algorithm 2 with the* LMC *critic from Algorithm 4. If we follow the hyperparameter choices of Lemma 14, then for any $(s,a) \in \mathcal{S}\times\mathcal{A}$,*

$$\left|\left\langle \phi(s,a), \left(\mu_h^{t,J_t} - \widehat{w}_h^t\right)\right\rangle\right| \le \frac{4}{3}\, \|\phi(s,a)\|_{(\Lambda_h^t)^{-1}}\,. \tag{C.94}$$

Lemma 17. *Consider Algorithm 2 with the* LMC *critic from Algorithm 4. If we follow the hyperparameter choices of Lemma 14, then for any $(s,a) \in \mathcal{S}\times\mathcal{A}$,*

$$\frac{1}{2\sqrt{6\,\zeta}}\, \|\phi(s,a)\|_{(\Lambda_h^t)^{-1}} \le \|\phi(s,a)\|_{\Sigma_h^{t,m,J_t}} \le \frac{4}{3}\sqrt{\frac{2}{3\,\zeta}}\, \|\phi(s,a)\|_{(\Lambda_h^t)^{-1}}\,. \tag{C.95}$$

Additionally, we outline the necessary supporting lemmas that are useful for bounding the model prediction error. Recall that $\left|\mathcal{D}^t\right| \coloneqq \sup_{h\in[H]} \left|\mathcal{D}_h^t\right|$ represents the number of trajectories in $\mathcal{D}^t$ or the number of (s_h, a_h, s_{h+1}) tuples in $\mathcal{D}_h^t$, where $\left|\mathcal{D}^t\right| = N$ in the on-policy setting, and $\left|\mathcal{D}^t\right| = t$ in the off-policy setting.

Lemma 18. *Consider Algorithm 2 with the* LMC *critic from Algorithm 4. For any $(t,h) \in [T]\times[H]$, it holds that*

$$\left\|\widehat{w}_h^t\right\|_2 \le 2\,H\,\sqrt{d_{\boldsymbol{c}}\,\left|\mathcal{D}^t\right|/\lambda}\,. \tag{C.96}$$

Lemma 19. *Consider Algorithm 2 with the* LMC *critic from Algorithm 4. If we follow the hyperparameter choices of Lemma 14, then for any $(t,m,h) \in [T]\times[M]\times[H]$ and for any $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\left\|w_h^{t,m,J_t}\right\|_2 \le \overline{W}_\delta^t \coloneqq \frac{16}{3}\, H\, \sqrt{d_{\boldsymbol{c}}\,\left|\mathcal{D}^t\right|} + \sqrt{\frac{2\,d_{\boldsymbol{c}}^3\,t}{3\,\zeta\,\delta}}\,. \tag{C.97}$$

Lemma 20. *Consider Algorithm 2 with the* LMC *critic from Algorithm 4. If we follow the hyperparameter choices of Lemma 14, then for any* $(t, m, h, s, a) \in [T] \times [M] \times [H] \times \mathcal{S} \times \mathcal{A}$ *and for any* $\delta \in (0, 1)$*, with probability at least* $1 - \delta$*,*

$$\left|\left\langle \phi(s,a), \widehat{w}_h^t - w_h^{t,m,J_t} \right\rangle\right| \leq \left(\frac{8}{3}\sqrt{\frac{2\, d_{\boldsymbol{c}}\, \log(1/\delta)}{3\,\zeta}} + \frac{4}{3}\right) \|\phi(s,a)\|_{(\Lambda_h^t)^{-1}} \,. \tag{C.98}$$

Lemma 21. *Consider Algorithm 2 with the* LMC *critic from Algorithm 4. Conditioned on* $\mathcal{E}_\delta$ *defined in Lemma 13, if we follow the hyperparameter choices of Lemma 14, then for any* $(t, h, s, a) \in [T] \times [H] \times \mathcal{S} \times \mathcal{A}$ *and for any* $\delta \in (0, 1)$*, it holds that*

$$\left|\left\langle \phi(s,a), \widehat{w}_h^t \right\rangle - r_h(s,a) - \mathbb{P}_h \widehat{V}_{h+1}^t(s,a)\right| \leq 3\, C_\delta\, H\, \sqrt{d_{\boldsymbol{c}}}\, \|\phi(s,a)\|_{(\Lambda_h^t)^{-1}} \,. \tag{C.99}$$

Main Analysis

We will use the above lemmas to complete the main proof in this section.

Proof of Lemma 14. **Optimism (RHS of Equation (C.89))** Using the definition of the model prediction error, we need to show that with high probability, $\widehat{Q}_h^t(s,a) \geq r_h(s,a) + \mathbb{P}_h \widehat{V}_{h+1}^t(s,a)$. Recall that $\widehat{Q}_h^t(s,a) = \min\left\{\left\langle \phi(s,a), w_h^{t,m,J_t}\right\rangle, H - h + 1\right\}$. Since $r_h(s,a) + \mathbb{P}_h \widehat{V}_{h+1}^t(s,a) \leq H - h + 1$, when $\left\langle \phi(s,a), w_h^{t,m,J_t}\right\rangle > H - h + 1$, the statement is trivially true. Thus, we only need to consider the case when $\left\langle \phi(s,a), w_h^{t,m,J_t}\right\rangle \leq H - h + 1$ and thus $\widehat{Q}_h^t(s,a) = \left\langle \phi(s,a), w_h^{t,m,J_t}\right\rangle$.

Based on the mean and covariance matrix defined in Lemma 15, we have that $\left\langle \phi(s,a), w_h^{t,m,J_t}\right\rangle$ follows the distribution $\mathbf{N}\left(\phi(s,a)^\top \mu_h^{t,J_t}, \phi(s,a)^\top\, \Sigma_h^{t,J_t}\, \phi(s,a)\right)$.

In order to prove that $\widehat{Q}_h^t(s,a) \geq r_h(s,a) + \mathbb{P}_h \widehat{V}_{h+1}^t(s,a)$, we consider the following variable $X_t \coloneqq \frac{r_h(s,a) + \mathbb{P}_h \widehat{V}_{h+1}^t(s,a) - \left\langle \phi(s,a), \mu_h^{t,J_t}\right\rangle}{\sqrt{\phi(s,a)^\top\, \Sigma_h^{t,J_t}\, \phi(s,a)}}$ and will next show that $|X_t| \leq 1$. First, we have that

$$\begin{aligned}
&\left|r_h(s,a) + \mathbb{P}_h \widehat{V}_{h+1}^t(s,a) - \left\langle \phi(s,a), \mu_h^{t,J_t}\right\rangle\right| && \text{(C.100)}\\
&\overset{\text{(i)}}{\leq} \left|r_h(s,a) + \mathbb{P}_h \widehat{V}_{h+1}^t(s,a) - \left\langle \phi(s,a), \widehat{w}_h^t\right\rangle\right| + \left|\left\langle \phi(s,a), \widehat{w}_h^t - \mu_h^{t,J_t}\right\rangle\right| && \text{(C.101)}\\
&\overset{\text{(ii)}}{\leq} C_\delta\, H\, \sqrt{d_{\boldsymbol{c}}}\, \|\phi(s,a)\|_{(\Lambda_h^t)^{-1}} + \frac{4}{3}\, \|\phi(s,a)\|_{(\Lambda_h^t)^{-1}} && \text{(C.102)}\\
&= \left(C_\delta\, H\, \sqrt{d_{\boldsymbol{c}}} + \frac{4}{3}\right) \|\phi(s,a)\|_{(\Lambda_h^t)^{-1}} \,, && \text{(C.103)}
\end{aligned}$$

where (i) uses the triangular inequality, and (ii) is implied by Lemmas 16 and 21. Therefore,

$$\begin{aligned}
|X_t| &= \left|\frac{r_h(s,a) + \mathbb{P}_h \widehat{V}_{h+1}^t(s,a) - \left\langle \phi(s,a), \mu_h^{t,J_t}\right\rangle}{\sqrt{\phi(s,a)^\top\, \Sigma_h^{t,J_t}\, \phi(s,a)}}\right| && \text{(C.104)}\\
&\leq \sqrt{\zeta}\left(2\, H\, \sqrt{d_{\boldsymbol{c}}} C_\delta + 8/3\right) \,. && \text{(C.105)}
\end{aligned}$$

Since we choose $\zeta = \left(2\,H\,\sqrt{d_c}C_\delta + 8/3\right)^{-2}$, we have that $|X_t| \leq 1$. Then, using Lemma 24, we can get that

$$\Pr\left(\left\langle \phi(s,a), w_h^{t,m,J_t}\right\rangle \geq r_h(s,a) + \mathbb{P}_h \widehat{V}_{h+1}^t(s,a)\right) \tag{C.106}$$

$$= \Pr\left(\frac{\left\langle \phi(s,a), w_h^{t,m,J_t}\right\rangle - \left\langle \phi(s,a), \mu_h^{t,J_t}\right\rangle}{\sqrt{\phi(s,a)^\top \Sigma_h^{t,J_t}\,\phi(s,a)}} \geq \frac{r_h(s,a) + \mathbb{P}_h \widehat{V}_{h+1}^t(s,a) - \left\langle \phi(s,a), \mu_h^{t,J_t}\right\rangle}{\sqrt{\phi(s,a)^\top \Sigma_h^{t,J_t}\,\phi(s,a)}}\right) \tag{C.107}$$

$$= \Pr\left(\frac{\left\langle \phi(s,a), w_h^{t,m,J_t}\right\rangle - \left\langle \phi(s,a), \mu_h^{t,J_t}\right\rangle}{\sqrt{\phi(s,a)^\top \Sigma_h^{t,J_t}\,\phi(s,a)}} \geq X_t\right) \tag{C.108}$$

$$\geq \frac{1}{2\sqrt{2\pi}} \exp\left(-X_t^2/2\right) \tag{C.109}$$

$$\geq \frac{1}{2\sqrt{2e\pi}}\,. \tag{C.110}$$

The above result holds for any $m \in [M]$. Since we have M parallel critic parameters, it holds that

$$\Pr\left(\exists (s,a) \in \mathcal{S}\times\mathcal{A} : \widehat{Q}_h^t(s,a) \leq r_h(s,a) + \mathbb{P}_h \widehat{V}_{h+1}^t(s,a)\right) \tag{C.111}$$

$$= \Pr\left(\exists (s,a) \in \mathcal{S}\times\mathcal{A} : \max_{m\in[M]} \widehat{Q}_h^{t,m}(s,a) \leq r_h(s,a) + \mathbb{P}_h \widehat{V}_{h+1}^t(s,a)\right) \tag{C.112}$$

$$= \Pr\left(\exists (s,a) \in \mathcal{S}\times\mathcal{A} : \forall m \in [M],\ \ \widehat{Q}_h^{t,m}(s,a) \leq r_h(s,a) + \mathbb{P}_h \widehat{V}_{h+1}^t(s,a)\right) \tag{C.113}$$

$$\leq \Pr\left(\forall m \in [M],\ \ \exists (s^m,a^m) \in \mathcal{S}\times\mathcal{A} : \widehat{Q}_h^{t,m}(s^m,a^m) \leq r_h(s^m,a^m) + \mathbb{P}_h \widehat{V}_{h+1}^t(s^m,a^m)\right) \tag{C.114}$$

$$= \prod_{m=1}^{M} \Pr\left(\exists (s,a) \in \mathcal{S}\times\mathcal{A} : \widehat{Q}_h^{t,m}(s,a) \leq r_h(s,a) + \mathbb{P}_h \widehat{V}_{h+1}^t(s,a)\right) \tag{C.115}$$

$$= \prod_{m=1}^{M} \left(1 - \Pr\left(\forall (s,a) \in \mathcal{S}\times\mathcal{A} : \widehat{Q}_h^{t,m}(s,a) \geq r_h(s,a) + \mathbb{P}_h \widehat{V}_{h+1}^t(s,a)\right)\right) \tag{C.116}$$

$$= \prod_{m=1}^{M} \left(1 - \Pr\left(\forall (s,a) \in \mathcal{S}\times\mathcal{A} : \left\langle \phi(s,a), w_h^{t,m,J_t}\right\rangle \geq r_h(s,a) + \mathbb{P}_h \widehat{V}_{h+1}^t(s,a)\right)\right) \tag{C.117}$$

$$\leq \left(1 - \frac{1}{2\sqrt{2e\pi}}\right)^M\,. \tag{C.118}$$

This further implies that

$$\Pr\left(\forall (s,a) \in \mathcal{S}\times\mathcal{A} : \iota_h^t(s,a) \leq 0\right) \tag{C.119}$$

$$= \Pr\left(\forall (s,a) \in \mathcal{S}\times\mathcal{A} : \widehat{Q}_h^t(s,a) \geq r_h(s,a) + \mathbb{P}_h \widehat{V}_{h+1}^t(s,a)\right) \tag{C.120}$$

$$= 1 - \Pr\left(\exists (s,a) \in \mathcal{S}\times\mathcal{A} : \widehat{Q}_h^t(s,a) \leq r_h(s,a) + \mathbb{P}_h \widehat{V}_{h+1}^t(s,a)\right) \tag{C.121}$$

$$= 1 - \left(1 - \frac{1}{2\sqrt{2e\pi}}\right)^M . \tag{C.122}$$

Let $1 - \left(1 - \frac{1}{2\sqrt{2e\pi}}\right)^M \geq 1 - \delta/(H\,T)$, which yields that $M = \log\left(H\,T/\delta\right)/\log\left(1/(1-c)\right)$ where $c = 1/(2\sqrt{2e\pi})$. Therefore, we have that

$$\Pr\left(\iota_h^t(s,a) \leq 0,\ \ \forall (s,a) \in \mathcal{S} \times \mathcal{A}\right) \geq 1 - \frac{\delta}{H\,T}\,. \tag{C.123}$$

Applying union bound over $[H]$ and $[T]$, we have that $\iota_h^t(s,a) \leq 0$ with probability $1-\delta$.

Error Bound (LHS of Equation (C.89)) We can lower bound ι_h^t as follows.

$$-\iota_h^t(s,a) = \widehat{Q}_h^t(s,a) - r_h(s,a) - \mathbb{P}_h \widehat{V}_{h+1}^t(s,a) \tag{C.124}$$

$$= \min\left\{\max_{m\in[M]} \left\langle \phi(s,a), w_h^{t,m,J_t}\right\rangle, H-h+1\right\}^+ - r_h(s,a) - \mathbb{P}_h \widehat{V}_{h+1}^t(s,a) \tag{C.125}$$

$$\leq \max_{m\in[M]} \left\langle \phi(s,a), w_h^{t,m,J_t}\right\rangle - r_h(s,a) - \mathbb{P}_h \widehat{V}_{h+1}^t(s,a) \tag{C.126}$$

$$= \max_{m\in[M]} \left\langle \phi(s,a), w_h^{t,m,J_t}\right\rangle - \left\langle \phi(s,a), \widehat{w}_h^t\right\rangle + \left\langle \phi(s,a), \widehat{w}_h^t\right\rangle - r_h(s,a) - \mathbb{P}_h \widehat{V}_{h+1}^t(s,a) \tag{C.127}$$

$$\leq \left|\max_{m\in[M]} \left\langle \phi(s,a), w_h^{t,m,J_t}\right\rangle - \left\langle \phi(s,a), \widehat{w}_h^t\right\rangle\right| + \left|\left\langle \phi(s,a), \widehat{w}_h^t\right\rangle - r_h(s,a) - \mathbb{P}_h \widehat{V}_{h+1}^t(s,a)\right| \tag{C.128}$$

$$\overset{\text{(iii)}}{\leq} \left(C_\delta\, H\, \sqrt{d_{\boldsymbol{c}}} + \frac{4}{3}\sqrt{\frac{2\,d_{\boldsymbol{c}} \log\left(1/\delta\right)}{3\,\zeta}} + \frac{4}{3}\right) \left\|\phi(s,a)\right\|_{(\Lambda_h^t)^{-1}}\,, \tag{C.129}$$

where (iii) is derived from Lemmas 20 and 21. This concludes the proof. □

C.4.2 Proofs of Preliminary Properties

Proof of Lemma 15

Proof. For any $(t,m) \in [T] \times [M]$, the critic update rule at j-th round can be written as

$$w_h^{t,m,j} = w_h^{t,m,j-1} - \alpha_{\boldsymbol{c}}^{h,t}\, \nabla L_h^t(w_h^{t,m,j-1}) + \sqrt{\alpha_{\boldsymbol{c}}^{h,t}\, \zeta^{-1}}\, \nu_h^{t,m,j}\,. \tag{C.130}$$

Considering $j = J_t$ and plugging in Equation (C.90), we have that

$$w_h^{t,m,J_t} = w_h^{t,m,J_t-1} - \alpha_{\boldsymbol{c}}^{h,t}\left(\Lambda_h^t\, w_h^{t,m,J_t-1} - b_h^t\right) + \sqrt{\alpha_{\boldsymbol{c}}^{h,t}\, \zeta^{-1}}\, \nu_h^{t,m,J_t} \tag{C.131}$$

$$= \left(I - \alpha_{\boldsymbol{c}}^{h,t}\, \Lambda_h^t\right)\, w_h^{t,m,J_t-1} + \alpha_{\boldsymbol{c}}^{h,t}\, b_h^t + \sqrt{\alpha_{\boldsymbol{c}}^{h,t}\, \zeta^{-1}}\, \nu_h^{t,m,J_t} \tag{C.132}$$

$$\overset{\text{(i)}}{=} \left(I - \alpha_{\boldsymbol{c}}^{h,t}\, \Lambda_h^t\right)^{J_t}\, w_h^{t,m,0} + \sum_{l=0}^{J_t-1} \left(I - \alpha_{\boldsymbol{c}}^{h,t}\, \Lambda_h^t\right)^l \left(\alpha_{\boldsymbol{c}}^{h,t} b_h^t + \sqrt{\alpha_{\boldsymbol{c}}^{h,t}\, \zeta^{-1}}\, \nu_h^{t,m,J_t-l}\right) \tag{C.133}$$

$$= \left(I - \alpha_{\boldsymbol{c}}^{h,t}\,\Lambda_h^t\right)^{J_t}\; w_h^{t,m,0} + \alpha_{\boldsymbol{c}}^{h,t}\sum_{l=0}^{J_t-1}\left(I - \alpha_{\boldsymbol{c}}^{h,t}\,\Lambda_h^t\right)^{l}\; b_h^t + \sqrt{\alpha_{\boldsymbol{c}}^{h,t}\,\zeta^{-1}}\sum_{l=0}^{J_t-1}\left(I - \alpha_{\boldsymbol{c}}^{h,t}\,\Lambda_h^t\right)^{l}\; \nu_h^{t,m,J_t-l} \tag{C.134}$$

$$\overset{\text{(ii)}}{=} A_t^{J_t}\, w_h^{t,m,0} + \alpha_{\boldsymbol{c}}^{h,t}\sum_{l=0}^{J_t-1} A_t^l\,\Lambda_h^t\,\widehat{w}_h^t + \sqrt{\alpha_{\boldsymbol{c}}^{h,t}\,\zeta^{-1}}\sum_{l=0}^{J_t-1} A_t^l\,\nu_h^{t,m,J_t-l} \tag{C.135}$$

$$\overset{\text{(iii)}}{=} A_t^{J_t}\, w_h^{t,m,0} + (I - A_t)\,\left(A_t^0 + A_t^1 + \ldots + A_t^{J_t-1}\right)\,\widehat{w}_h^t + \sqrt{\alpha_{\boldsymbol{c}}^{h,t}\,\zeta^{-1}}\sum_{l=0}^{J_t-1} A_t^l\nu_h^{t,m,J_t-l} \tag{C.136}$$

$$\overset{\text{(iv)}}{=} A_t^{J_t}\, w_h^{t,m,0} + \left(I - A_t^{J_t}\right)\,\widehat{w}_h^t + \sqrt{\alpha_{\boldsymbol{c}}^{h,t}\,\zeta^{-1}}\sum_{l=0}^{J_t-1} A_t^l\,\nu_h^{t,m,J_t-l}\,. \tag{C.137}$$

(i) comes from telescoping the previous equation from $l = 0$ to $J_t - 1$. (ii) uses the definition that $A_t = I - \alpha_{\boldsymbol{c}}^{h,t}\,\Lambda_h^t$ and $b_h^t = \Lambda_h^t\widehat{w}_h^t$. (iii) uses the definition of A_t. (iv) follows from $I + A + \ldots + A^{n-1} = (I - A^n)(I - A)^{-1}$. Since we set $\alpha_{\boldsymbol{c}}^{h,t} = 1/\left(2\,\lambda_{\max}(\Lambda_h^t)\right.$, A_t satisfies $I \succ A_t \succ 0$ for all $t \in [T]$. Note that we warm-start the parameters from the previous episode and set $w_h^{t,m,0} = w_h^{t-1,m,J_{t-1}}$. Therefore, by telescoping the above equation from $i = 0$ to t, we further have that

$$w_h^{t,m,J_t} = A_t^{J_t}\, w_h^{t-1,m,J_{t-1}} + \left(I - A_t^{J_t}\right)\,\widehat{w}_h^t + \sqrt{\alpha_{\boldsymbol{c}}^{h,t}\,\zeta^{-1}}\sum_{l=0}^{J_t-1} A_t^l\,\nu_h^{t,m,J_t-l} \tag{C.138}$$

$$= A_t^{J_t}\ldots A_1^{J_1} w_h^{1,m,0} + \sum_{i=1}^{t} A_t^{J_t}\ldots A_{i+1}^{J_{i+1}}\,\left(I - A_i^{J_i}\right)\,\widehat{w}_h^i + \sum_{i=1}^{t}\sqrt{\alpha_{\boldsymbol{c}}^i\,\zeta^{-1}}A_t^{J_t}\ldots A_{i+1}^{J_{i+1}}\sum_{l=0}^{J_i-1} A_i^l\nu_h^{i,J_i-l}\,. \tag{C.139}$$

Note that if $\xi \sim \mathbf{N}(0, I_{d\times d})$, then we have that $A\,\xi + \mu \sim \mathbf{N}(\mu, A\,A^\top)$ for any $A \in \mathbb{R}^{d\times d}$ and $\mu \in \mathbb{R}^d$. This implies that w_h^{t,m,J_t} follows the Gaussian distribution $N(\mu_h^{t,m,J_t}, \Sigma_h^{t,m,J_t})$, where

$$\mu_h^{t,m,J_t} = A_t^{J_t}\ldots A_1^{J_1} w_h^{1,m,0} + \sum_{i=1}^{t} A_t^{J_t}\ldots A_{i+1}^{J_{i+1}}\,\left(I - A_i^{J_i}\right)\,\widehat{w}_h^i\,. \tag{C.140}$$

We then derive the covariance matrix Σ_h^{t,m,J_t}. For any $i \in [t]$, we denote that $\mathscr{A}_{i+1} = A_t^{J_t}\ldots A_{i+1}^{J_{i+1}}$. Therefore,

$$\sqrt{\alpha_{\boldsymbol{c}}^i\,\zeta^{-1}}\,\mathscr{A}_{i+1}\sum_{l=0}^{J_i-1} A_i^l\,\nu_h^{i,J_i-l} = \sum_{l=0}^{J_i-1}\sqrt{\alpha_{\boldsymbol{c}}^i\,\zeta^{-1}}\,\mathscr{A}_{i+1}\,A_i^l\,\nu_h^{i,J_i-l} \tag{C.141}$$

$$\sim \mathbf{N}\left(0,\,\sum_{l=0}^{J_i-1}\alpha_{\boldsymbol{c}}^i\,\zeta^{-1}\,\mathscr{A}_{i+1}\,A_i^l\,(\mathscr{A}_{i+1}\,A_i^l)^\top\right) \sim \mathbf{N}\left(0, \alpha_{\boldsymbol{c}}^i\,\zeta^{-1}\,\mathscr{A}_{i+1}\,\left(\sum_{l=0}^{J_i-1} A_i^{2l}\right)\,\mathscr{A}_{i+1}^\top\right)\,. \tag{C.142}$$

This further implies that

$$\Sigma_h^{t,m,J_t} = \sum_{i=1}^{t} \alpha_{\boldsymbol{c}}^i \, \zeta^{-1} \, \mathscr{A}_{i+1} \left(\sum_{l=0}^{J_i-1} A_i^{2l} \right) \mathscr{A}_{i+1}^\top \tag{C.143}$$

$$= \sum_{i=1}^{t} \alpha_{\boldsymbol{c}}^i \, \zeta^{-1} \, A_t^{J_t} \dots A_{i+1}^{J_{i+1}} \left(\sum_{l=0}^{J_i-1} A_i^{2l} \right) A_{i+1}^{J_{i+1}} \dots A_t^{J_t} \tag{C.144}$$

$$\overset{\text{(v)}}{=} \sum_{i=1}^{t} \alpha_{\boldsymbol{c}}^i \, \zeta^{-1} \, A_t^{J_t} \dots A_{i+1}^{J_{i+1}} \left(I - A_i^{2J_i} \right) \left(I - A_i^2 \right)^{-1} A_{i+1}^{J_{i+1}} \dots A_t^{J_t} \tag{C.145}$$

$$= \sum_{i=1}^{t} \alpha_{\boldsymbol{c}}^i \, \zeta^{-1} \, A_t^{J_t} \dots A_{i+1}^{J_{i+1}} \left(I - A_i^{2J_i} \right) \left(\Lambda_h^i \right) \left(\Lambda_h^i \right)^{-1} (I - A_i)^{-1} \, (I + A_i)^{-1} \, A_{i+1}^{J_{i+1}} \dots A_t^{J_t} \tag{C.146}$$

$$\overset{\text{(vi)}}{=} \sum_{i=1}^{t} \zeta^{-1} \, A_t^{J_t} \dots A_{i+1}^{J_{i+1}} \left(I - A_i^{2J_i} \right) \left(\Lambda_h^i \right)^{-1} \, (I + A_i)^{-1} \, A_{i+1}^{J_{i+1}} \dots A_t^{J_t} \, . \tag{C.147}$$

(v) uses the fact that $I + A + \ldots + A^{n-1} = (I - A^n)(I - A)^{-1}$, and (vi) uses the fact that $\alpha_{\boldsymbol{c}}^{h,t} \, \Lambda_h^t = I - A_t$. This concludes the proof. □

Proof of Lemma 16

Proof. Using Lemma 15, we first have that

$$\mu_h^{t,J_t} = A_t^{J_t} \dots A_1^{J_1} w_h^{1,m,0} + \sum_{i=1}^{t} A_t^{J_t} \dots A_{i+1}^{J_{i+1}} \left(I - A_i^{J_i} \right) \widehat{w}_h^i \tag{C.148}$$

$$= A_t^{J_t} \dots A_1^{J_1} w_h^{1,m,0} + \sum_{i=1}^{t} A_t^{J_t} \dots A_{i+1}^{J_{i+1}} \, \widehat{w}_h^i - \sum_{i=1}^{t} A_t^{J_t} \dots A_i^{J_i} \, \widehat{w}_h^i \tag{C.149}$$

$$= A_t^{J_t} \dots A_1^{J_1} w_h^{1,m,0} + \sum_{i=1}^{t-1} A_t^{J_t} \dots A_{i+1}^{J_{i+1}} \left(\widehat{w}_h^i - \widehat{w}_h^{i+1} \right) - A_t^{J_t} \dots A_1^{J_1} \, \widehat{w}_h^1 + \widehat{w}_h^t \tag{C.150}$$

$$= A_t^{J_t} \dots A_1^{J_1} \left(w_h^{1,m,0} - \widehat{w}_h^1 \right) + \sum_{i=1}^{t-1} A_t^{J_t} \dots A_{i+1}^{J_{i+1}} \left(\widehat{w}_h^i - \widehat{w}_h^{i+1} \right) + \widehat{w}_h^t \, . \tag{C.151}$$

This implies that

$$\left| \left\langle \phi(s,a), \left(\mu_h^{t,J_t} - \widehat{w}_h^t \right) \right\rangle \right| = \phi(s,a)^\top A_t^{J_t} \dots A_1^{J_1} \left(w_h^{1,m,0} - \widehat{w}_h^1 \right) + \phi(s,a)^\top \sum_{i=1}^{t-1} A_t^{J_t} \dots A_{i+1}^{J_{i+1}} \left(\widehat{w}_h^i - \widehat{w}_h^{i+1} \right) \tag{C.152}$$

$$\overset{\text{(i)}}{=} \left| \phi(s,a)^\top \sum_{i=0}^{t-1} A_t^{J_t} \dots A_{i+1}^{J_{i+1}} \left(\widehat{w}_h^i - \widehat{w}_h^{i+1} \right) \right| \tag{C.153}$$

$$= \left| \sum_{i=0}^{t-1} \phi(s,a)^\top A_t^{J_t} \dots A_{i+1}^{J_{i+1}} \left(\widehat{w}_h^i - \widehat{w}_h^{i+1} \right) \right| \tag{C.154}$$

$$\overset{\text{(ii)}}{\leq} \sum_{i=0}^{t-1} \prod_{j=i+1}^{t} \left(1 - \alpha_{\boldsymbol{c}}^{h,j}\, \lambda_{\min}\left(\Lambda_h^j\right)\right)^{J_j} \|\phi(s,a)\|_2\, \|\widehat{w}_h^i - \widehat{w}_h^{i+1}\|_2 \tag{C.155}$$

$$\overset{\text{(iii)}}{\leq} \sum_{i=0}^{t-1} \prod_{j=i+1}^{t} \left(1 - \alpha_{\boldsymbol{c}}^{h,j}\, \lambda_{\min}\left(\Lambda_h^j\right)\right)^{J_j} \|\phi(s,a)\|_2 \left(\|\widehat{w}_h^i\|_2 + \|\widehat{w}_h^{i+1}\|_2\right) \tag{C.156}$$

$$\overset{\text{(iv)}}{\leq} 4\,H\,\sqrt{d_{\boldsymbol{c}}\,|\mathcal{D}^t|\,/\lambda} \sum_{i=0}^{t-1} \prod_{j=i+1}^{t} \left(1 - \alpha_{\boldsymbol{c}}^{h,j}\, \lambda_{\min}\left(\Lambda_h^j\right)\right)^{J_j} \|\phi(s,a)\|_2 \tag{C.157}$$

$$\overset{\text{(v)}}{\leq} 4\,H\,\left(\left|\mathcal{D}^t\right| + 1\right)\,\sqrt{d_{\boldsymbol{c}}/\lambda} \sum_{i=0}^{t-1} \prod_{j=i+1}^{t} \left(1 - \alpha_{\boldsymbol{c}}^{h,j}\, \lambda_{\min}\left(\Lambda_h^j\right)\right)^{J_j} \|\phi(s,a)\|_{\left(\Lambda_h^i\right)^{-1}} \tag{C.158}$$

$$\overset{\text{(vi)}}{\leq} 4\,H\,\left(\left|\mathcal{D}^t\right| + 1\right)\,\sqrt{d_{\boldsymbol{c}}} \sum_{i=0}^{t-1} \sigma^{t-i}\, \|\phi(s,a)\|_{\left(\Lambda_h^i\right)^{-1}} \tag{C.159}$$

$$\overset{\text{(vii)}}{\leq} 4\,H\,\left(\left|\mathcal{D}^t\right| + 1\right)\,\sqrt{d_{\boldsymbol{c}}} \left(\sum_{i=0}^{t-1} \sigma^{t-i}\right) \|\phi(s,a)\|_{\left(\Lambda_h^t\right)^{-1}} \tag{C.160}$$

$$\leq 4\,H\,\left(\left|\mathcal{D}^t\right| + 1\right)\,\sqrt{d_{\boldsymbol{c}}} \left(\sum_{i=0}^{t-1} \sigma^{t-i}\right) \|\phi(s,a)\|_{\left(\Lambda_h^t\right)^{-1}} \tag{C.161}$$

$$= 4\,H\,\left(\left|\mathcal{D}^t\right| + 1\right)\,\sqrt{d_{\boldsymbol{c}}} \left(\sum_{i=1}^{t-1} \sigma^{i}\right) \|\phi(s,a)\|_{\left(\Lambda_h^t\right)^{-1}} \tag{C.162}$$

$$\overset{\text{(viii)}}{\leq} 4\,H\,\left(\left|\mathcal{D}^t\right| + 1\right)\,\sqrt{d_{\boldsymbol{c}}} \left(\frac{\sigma}{1-\sigma}\right) \|\phi(s,a)\|_{\left(\Lambda_h^t\right)^{-1}} \tag{C.163}$$

$$= \left(\frac{1}{1-\sigma}\right) \|\phi(s,a)\|_{\left(\Lambda_h^t\right)^{-1}} \tag{C.164}$$

$$\leq \frac{4}{3}\, \|\phi(s,a)\|_{\left(\Lambda_h^t\right)^{-1}}\,. \tag{C.165}$$

For (i), we choose $w_h^{1,m,0} = \mathbf{0}$ and denote that $\widehat{w}_h^0 = \mathbf{0}$. (ii) comes from $A_i \prec (1 - \alpha_{\boldsymbol{c}}^{h,j}\, \lambda_{\min}\left(\Lambda_h^j\right)) I$ and the Hölder's inequality. (iii) uses the triangular inequality. (iv) uses Lemma 18. (v) uses the fact that $\|\phi(s,a)\| \leq \sqrt{|\mathcal{D}^t| + 1}\, \|\phi(s,a)\|_{(\Lambda_h^i)^{-1}}$. (vi) hold because we set $\lambda = 1$ and uses Lemma 28 by setting $J_j \geq \kappa_j \log(1/\sigma)$ where $\sigma = 1/\left(4\,H\,\left(\left|\mathcal{D}^t\right| + 1\right)\,\sqrt{d_{\boldsymbol{c}}}\right)$. (vii) follows from $\|\phi(s,a)\|_{\left(\Lambda_h^i\right)^{-1}} \leq \|\phi(s,a)\|_2 \leq \sqrt{|\mathcal{D}^t| + 1}\, \|\phi(s,a)\|_{\left(\Lambda_h^t\right)^{-1}}$. (viii) follows from $\sum_{i=1}^t \sigma^t \leq \sum_{i=1}^\infty \sigma^i \leq \sigma/(1-\sigma)$. This concludes the proof. □

Proof of Lemma 17

Proof. We first bound the RHS. Using Lemma 15, we have that

$$\phi(s,a)^\top\, \Sigma_h^{t,J_t}\, \phi(s,a) \tag{C.166}$$

$$= \frac{1}{\zeta} \sum_{i=1}^{t} \phi(s,a)^\top\, A_t^{J_t} \ldots A_{i+1}^{J_{i+1}} \left(I - A^{2J_i}\right) \left(\Lambda_h^i\right)^{-1} (I + A_i)^{-1}\, A_{i+1}^{J_{i+1}} \ldots A_t^{J_t}\, \phi(s,a) \tag{C.167}$$

$$\overset{\text{(i)}}{=} \frac{1}{\zeta}\sum_{i=1}^{t}\phi(s,a)^\top \mathscr{A}_{i+1}\left(I-A^{2J_i}\right)\left(\Lambda_h^i\right)^{-1}(I+A_i)^{-1}\mathscr{A}_{i+1}^\top\phi(s,a) \tag{C.168}$$

$$\overset{\text{(ii)}}{\leq} \frac{2}{3\,\zeta}\sum_{i=1}^{t}\phi(s,a)^\top\mathscr{A}_{i+1}\left(\left(\Lambda_h^i\right)^{-1}-A_i^{J_i}\left(\Lambda_h^i\right)^{-1}A_i^{J_i}\right)\mathscr{A}_{i+1}^\top\phi(s,a) \tag{C.169}$$

$$= \frac{2}{3\,\zeta}\left(\sum_{i=1}^{t}\phi(s,a)^\top\mathscr{A}_{i+1}\left(\Lambda_h^i\right)^{-1}\mathscr{A}_{i+1}^\top\phi(s,a)-\sum_{i=1}^{t}\phi(s,a)^\top\mathscr{A}_i\left(\Lambda_h^i\right)^{-1}\mathscr{A}_i^\top\phi(s,a)\right) \tag{C.170}$$

$$\overset{\text{(iii)}}{\leq} \frac{2}{3\,\zeta}\sum_{i=1}^{t}\phi(s,a)^\top\mathscr{A}_{i+1}\left(\Lambda_h^i\right)^{-1}\mathscr{A}_{i+1}^\top\phi(s,a) \tag{C.171}$$

$$= \frac{2}{3\,\zeta}\left(\|\phi(s,a)\|^2_{(\Lambda_h^i)^{-1}}+\sum_{i=1}^{t-1}\left\|\mathscr{A}_{i+1}^\top\phi(s,a)\right\|^2_{(\Lambda_h^i)^{-1}}\right) \tag{C.172}$$

$$\leq \frac{2}{3\,\zeta}\,\|\phi(s,a)\|^2_{(\Lambda_h^t)^{-1}}+\frac{2}{3\,\zeta}\sum_{i=1}^{t-1}\prod_{j=i+1}^{t}\left(1-\alpha_{\boldsymbol{c}}\,\lambda_{\min}(\Lambda_h^j)\right)^{2\,J_j}\|\phi(s,a)\|^2_{(\Lambda_h^i)^{-1}}\,. \tag{C.173}$$

For (i), we denote $\mathscr{A}_{i+1}=A_t^{J_t}\ldots A_{i+1}^{J_{i+1}}$. (ii) follows from $I+A_i\succeq\frac{3}{2}I$ since we set $\alpha_{\boldsymbol{c}}^{h,j}=1/\left(2\,\lambda_{\max}(\Lambda_h^j)\right)$. In particular, it is easy to prove that A and $(\Lambda_h^t)^{-1}$ are commuting matrices and thus

$$A^{2J_i}\left(\Lambda_h^i\right)^{-1} = A^{2J_i-1}\left(I-\alpha_{\boldsymbol{c}}^{h,t}\,\Lambda_h^t\right)(\Lambda_h^t)^{-1} \tag{C.174}$$

$$= A^{2J_i-1}(\Lambda_h^t)^{-1}\left(I-\alpha_{\boldsymbol{c}}^{h,t}\,\Lambda_h^t\right) \tag{C.175}$$

$$= A^{2J_i-1}(\Lambda_h^t)^{-1}A \tag{C.176}$$

$$\vdots \tag{C.177}$$

$$= A^{J_i}(\Lambda_h^t)^{-1}A^{J_i}\,. \tag{C.178}$$

(iii) follows from the fact that $\sum_{i=1}^{t}\phi(s,a)^\top\mathscr{A}_i\left(\Lambda_h^i\right)^{-1}\mathscr{A}_i^\top\phi(s,a)>0$. Therefore,

$$\left\|\phi(s,a)^\top\left(\Sigma_h^{t,J_t}\right)^{1/2}\right\|_2=\sqrt{\phi(s,a)^\top\,\Sigma_h^{t,J_t}\,\phi(s,a)} \tag{C.179}$$

$$\overset{\text{(iv)}}{\leq}\sqrt{\frac{2}{3\,\zeta}}\,\|\phi(s,a)\|_{(\Lambda_h^t)^{-1}}+\sqrt{\frac{2}{3\,\zeta}}\sum_{i=1}^{t-1}\prod_{j=i+1}^{t}\left(1-\alpha_{\boldsymbol{c}}\,\lambda_{\min}(\Lambda_h^j)\right)^{J_j}\|\phi(s,a)\|_{(\Lambda_h^i)^{-1}} \tag{C.180}$$

$$\overset{\text{(v)}}{\leq}\sqrt{\frac{2}{3\,\zeta}}\,\|\phi(s,a)\|_{(\Lambda_h^t)^{-1}}+\sqrt{\frac{2}{3\,\zeta}}\sum_{i=1}^{t-1}\sigma^{t-i}\,\|\phi(s,a)\|_{(\Lambda_h^i)^{-1}} \tag{C.181}$$

$$\overset{\text{(vi)}}{\leq}\sqrt{\frac{2}{3\,\zeta}}\,\|\phi(s,a)\|_{(\Lambda_h^t)^{-1}}+\sqrt{\frac{2\,(|\mathcal{D}^t|+1)}{3\,\zeta}}\left(\sum_{i=1}^{t-1}\sigma^{t-i}\right)\|\phi(s,a)\|_{(\Lambda_h^t)^{-1}} \tag{C.182}$$

$$\overset{\text{(vii)}}{\leq}\sqrt{\frac{2}{3\,\zeta}}\,\|\phi(s,a)\|_{(\Lambda_h^t)^{-1}}+\sqrt{\frac{2\,(|\mathcal{D}^t|+1)}{3\,\zeta}}\left(\frac{\sigma}{1-\sigma}\right)\|\phi(s,a)\|_{(\Lambda_h^t)^{-1}} \tag{C.183}$$

$$\leq\sqrt{\frac{2}{3\,\zeta}}\,\|\phi(s,a)\|_{(\Lambda_h^t)^{-1}}+\frac{1}{4}\sqrt{\frac{2}{3\,\zeta}}\left(\frac{1}{1-\sigma}\right)\|\phi(s,a)\|_{(\Lambda_h^t)^{-1}} \tag{C.184}$$

$$\leq \left(\sqrt{\frac{2}{3\,\zeta}} + \frac{1}{3}\sqrt{\frac{2}{3\,\zeta}}\right) \|\phi(s,a)\|_{(\Lambda_h^t)^{-1}} \tag{C.185}$$

$$\leq \frac{4}{3}\sqrt{\frac{2}{3\,\zeta}}\, \|\phi(s,a)\|_{(\Lambda_h^t)^{-1}}\ . \tag{C.186}$$

$$\tag{C.187}$$

(iv) follows from the fact that $\sqrt{a+b} \leq a+b$ for all $a, b > 0$. (v) uses Lemma 28 by setting $J_j \geq 2\,\kappa_j \log(1/\sigma)$ where $\sigma = 1/\left(4\,H\,\left(\left|\mathcal{D}^t\right|+1\right)\,\sqrt{d_{\boldsymbol{c}}}\right)$. (vi) follows from $\|\phi(s,a)\|_{\left(\Lambda_h^i\right)^{-1}} \leq \|\phi(s,a)\|_2 \leq \sqrt{|\mathcal{D}^t|+1}\,\|\phi(s,a)\|_{\left(\Lambda_h^t\right)^{-1}}$. (vii) follows from $\sum_{i=1}^t \sigma^{t-i} \leq \sum_{i=1}^\infty \sigma^i \leq \sigma/(1-\sigma)$.

We then proceed to bound the LHS. Using the definition of Σ_h^{t,J_t} from Equation (C.93), we have

$$\phi(s,a)^\top\,\Sigma_h^{t,J_t}\,\phi(s,a) \tag{C.188}$$

$$= \sum_{i=1}^t \frac{1}{\zeta}\phi(s,a)^\top A_t^{J_t}\ldots A_{i+1}^{J_{i+1}}\,\left(I - A^{2J_i}\right)\,\left(\Lambda_h^i\right)^{-1}\,(I+A_i)^{-1}\,A_{i+1}^{J_{i+1}}\ldots A_t^{J_t}\phi(s,a) \tag{C.189}$$

$$\overset{\text{(iii)}}{\geq} \frac{1}{2\,\zeta}\sum_{i=1}^t \phi(s,a)^\top \mathscr{A}_{i+1}\,\left(I - A^{2J_i}\right)\,\left(\Lambda_h^i\right)^{-1}\,\mathscr{A}_{i+1}^\top\,\phi(s,a) \tag{C.190}$$

$$= \frac{1}{2\,\zeta}\,\sum_{i=1}^t \frac{1}{2\,\zeta}\phi(s,a)^\top \mathscr{A}_{i+1}\,\left(\left(\Lambda_h^i\right)^{-1} - A_t^{J_t}\,\left(\Lambda_h^i\right)^{-1}\,A_t^{J_t}\right)\mathscr{A}_{i+1}^\top\,\phi(s,a) \tag{C.191}$$

$$= \frac{1}{2\,\zeta}\,\sum_{i=1}^{t-1} \phi(s,a)^\top \mathscr{A}_{i+1}\,\left((\Lambda_h^i)^{-1} - (\Lambda_h^{i+1})^{-1}\right)\mathscr{A}_{i+1}^\top\,\phi(s,a) \tag{C.192}$$

$$- \frac{1}{2\,\zeta}\phi(s,a)^\top A_t^{J_t}\ldots A_1^{J_1}(\Lambda_h^1)^{-1}A_1^{J_1}\ldots A_t^{J_t}\phi(s,a) + \frac{1}{2\,\zeta}\phi(s,a)^\top(\Lambda_h^t)^{-1}\phi(s,a)\,, \tag{C.193}$$

where (iii) follows from $(I+A_t)^{-1} \succeq \frac{1}{2}\,I$ for all $t \in [T]$.

$$\left|\phi(s,a)^\top \mathscr{A}_{i+1}\,\left((\Lambda_h^i)^{-1} - (\Lambda_h^{i+1})^{-1}\right)\mathscr{A}_{i+1}^\top\,\phi(s,a)\right| \tag{C.194}$$

$$\leq \left|\phi(s,a)^\top \mathscr{A}_{i+1}\,(\Lambda_h^i)^{-1}\,\mathscr{A}_{i+1}^\top\,\phi(s,a)\right| \tag{C.195}$$

$$+ \left|\left\langle \phi(s,a), \mathscr{A}_{i+1}\,(\Lambda_h^{i+1})^{-1}\,\mathscr{A}_{i+1}^\top\,\phi(s,a)\right\rangle\right| \tag{C.196}$$

$$\leq \left\|\phi(s,a)^\top \mathscr{A}_{i+1}\,(\Lambda_h^i)^{-1/2}\right\|^2 + \left\|\phi(s,a)^\top \mathscr{A}_{i+1}\,(\Lambda_h^{i+1})^{-1/2}\right\|^2 \tag{C.197}$$

$$= \prod_{j=i+1}^t \left(1 - \alpha_{\boldsymbol{c}}^{h,j}\,\lambda_{\min}(\Lambda_h^j)\right)^{2\,J_j}\,\left(\|\phi(s,a)\|^2_{(\Lambda_h^i)^{-1}} + \|\phi(s,a)\|^2_{(\Lambda_h^{i+1})^{-1}}\right) \tag{C.198}$$

$$\leq 2\,\prod_{j=i+1}^t \left(1 - \alpha_{\boldsymbol{c}}^{h,j}\,\lambda_{\min}(\Lambda_h^j)\right)^{2\,J_j}\,\|\phi(s,a)\|_2^2\,, \tag{C.199}$$

where we used $0 < \|\phi(s,a)\|_{(\Lambda_h^i)^{-1}} \leq \|\phi(s,a)\|_2$. Therefore, we have that

$$\phi(s,a)^\top \Sigma_h^{t,J_t}\, \phi(s,a) \tag{C.200}$$

$$\geq \frac{1}{2\,\zeta}\, \|\phi(s,a)\|^2_{(\Lambda_h^t)^{-1}} - \frac{1}{2\,\zeta} \prod_{i=1}^{t} \left(1 - \alpha_{\boldsymbol{c}}^{h,j} \lambda_{\min}(\Lambda_h^i)\right)^{2J_i} \|\phi(s,a)\|_2^2 \tag{C.201}$$

$$- \frac{1}{\zeta} \sum_{i=1}^{t-1} \prod_{j=i+1}^{t} \left(1 - \alpha_{\boldsymbol{c}}^{h,j}\, \lambda_{\min}(\Lambda_h^j)\right)^{2\,J_j} \|\phi(s,a)\|_2^2 \tag{C.202}$$

$$\overset{\text{(iv)}}{\geq} \frac{1}{2\,\zeta} \left(\|\phi(s,a)\|^2_{(\Lambda_h^t)^{-1}} - \sigma^t\, \|\phi(s,a)\|_2^2 - \sum_{i=1}^{t-1} 2\,\sigma^i\, \|\phi(s,a)\|_2^2 \right) \tag{C.203}$$

$$\overset{\text{(v)}}{\geq} \frac{1}{2\,\zeta}\, \|\phi(s,a)\|^2_{(\Lambda_h^t)^{-1}} \left(1 - \left(\left|\mathcal{D}^t\right| + 1\right) \sigma^t - 2\, \left(\left|\mathcal{D}^t\right| + 1\right) \sum_{i=1}^{t-1} \sigma^i \right) \tag{C.204}$$

$$\geq \frac{1}{2\,\zeta}\, \|\phi(s,a)\|^2_{(\Lambda_h^t)^{-1}} \left(1 - \sigma^{t-1} - \frac{1}{2\,(1-\sigma)} \right) \tag{C.205}$$

$$\geq \frac{1}{2\,\zeta}\, \|\phi(s,a)\|^2_{(\Lambda_h^t)^{-1}} \left(1 - \frac{1}{4} - \frac{2}{3} \right) \tag{C.206}$$

$$= \frac{1}{24\,\zeta}\, \|\phi(s,a)\|^2_{(\Lambda_h^t)^{-1}}\,, \tag{C.207}$$

where (iv) uses Lemma 28 by setting $J_j \geq 2\,\kappa_j \log(1/\sigma)$ where $\sigma = 1/\left(4\,H\,\left(\left|\mathcal{D}^t\right| + 1\right)\,\sqrt{d_{\boldsymbol{c}}}\right)$, and (v) use $\|\phi(s,a)\|_2 \leq \sqrt{|\mathcal{D}^t| + 1}\, \|\phi(s,a)\|_{(\Lambda_h^t)^{-1}}$. This concludes the proof. □

Proof of Lemma 18

Proof. Given the definition of $\widehat{w}_h^t$ in Equation (C.91), we have that

$$\left\|\widehat{w}_h^t\right\| = \left\| \left(\Lambda_h^t\right)^{-1} \sum_{(s,a)\in\mathcal{D}_h^t} \left[r_h(s,a) + \widehat{V}_{h+1}^t(s) \right] \cdot \phi(s_h, a) \right\| \tag{C.208}$$

$$\leq \sqrt{\frac{|\mathcal{D}^t|}{\lambda}} \left(\sum_{(s,a)\in\mathcal{D}_h^t} \left\| \left[r_h(s,a) + \widehat{V}_{h+1}^t(s) \right] \cdot \phi(s,a) \right\|^2_{(\Lambda_h^t)^{-1}} \right)^{1/2} \tag{C.209}$$

$$\leq 2\,H \sqrt{\frac{|\mathcal{D}^t|}{\lambda}} \left(\sum_{(s,a)\in\mathcal{D}_h^t} \|\phi(s,a)\|^2_{(\Lambda_h^t)^{-1}} \right)^{1/2} \tag{C.210}$$

$$\leq 2\,H \sqrt{d_{\boldsymbol{c}}\, |\mathcal{D}^t|\, /\lambda}\,, \tag{C.211}$$

where the first inequality follows from Lemma 27, the second inequality is due to the fact that $V_h^t \in [0, H]$ and the reward function is bounded by 1, and the last inequality follows from Lemma 22. □

Proof of Lemma 19

Proof. From Lemma 15, we know w_h^{t,m,J_t} follows Gaussian distribution $\mathbf{N}(\mu_h^{t,J_t}, \Sigma_h^{t,J_t})$. Therefore, we have that

$$\left\|w_h^{t,m,J_t}\right\|_2 = \left\|\mu_h^{t,J_t} + \xi_h^{t,J_t}\right\|_2 \leq \underbrace{\left\|\mu_h^{t,J_t}\right\|_2}_{\text{(I)}} + \underbrace{\left\|\xi_h^{t,J_t}\right\|_2}_{\text{(II)}}, \tag{C.212}$$

where $\xi_h^{t,J_t} \sim \mathbf{N}(0, \Sigma_h^{t,J_t})$. We first start by bounding Term (I). Given Lemma 15, by setting $w_h^{1,m,0} = \mathbf{0}$, we can obtain that

$$\left\|\mu_h^{t,J_t}\right\|_2 = \left\|A_t^{J_t} \dots A_1^{J_1} w_h^{1,m,0} + \sum_{i=1}^{t} A_t^{J_t} \dots A_{i+1}^{J_{i+1}} \left(I - A_i^{J_i}\right) \widehat{w}_h^i\right\|_2 \tag{C.213}$$

$$\leq \sum_{i=1}^{t} \left\|A_t^{J_t} \dots A_{i+1}^{J_{i+1}} \left(I - A_i^{J_i}\right) \widehat{w}_h^i\right\|_2 \tag{C.214}$$

$$\overset{\text{(i)}}{\leq} \sum_{i=1}^{t} \left\|A_t^{J_t} \dots A_{i+1}^{J_{i+1}} \left(I - A_i^{J_i}\right)\right\|_2 \left\|\widehat{w}_h^i\right\|_2 \tag{C.215}$$

$$\overset{\text{(ii)}}{\leq} 2\,H\,\sqrt{d_{\boldsymbol{c}}\,|\mathcal{D}^t|} \sum_{i=1}^{t} \left\|A_t^{J_t} \dots A_{i+1}^{J_{i+1}} \left(I - A_i^{J_i}\right)\right\|_2 \tag{C.216}$$

$$\overset{\text{(iii)}}{\leq} 2\,H\,\sqrt{d_{\boldsymbol{c}}\,|\mathcal{D}^t|} \sum_{i=1}^{t} \|A_t\|_2^{J_t} \dots \|A_{i+1}\|_2^{J_{i+1}} \left\|\left(I - A_i^{J_i}\right)\right\|_2 \tag{C.217}$$

$$\overset{\text{(iv)}}{\leq} 2\,H\,\sqrt{d_{\boldsymbol{c}}\,|\mathcal{D}^t|} \sum_{i=1}^{t} \prod_{j=i+1}^{t} \left(1 - \alpha_{\boldsymbol{c}}^{h,j}\,\lambda_{\min}(\Lambda_h^j)\right)^{J_j} \left(\|I\|_2 + \|A_i^{J_i}\|_2\right) \tag{C.218}$$

$$\overset{\text{(v)}}{\leq} 2\,H\,\sqrt{d_{\boldsymbol{c}}\,|\mathcal{D}^t|} \sum_{i=1}^{t} \prod_{j=i+1}^{t} \left(1 - \alpha_{\boldsymbol{c}}^{h,j}\,\lambda_{\min}(\Lambda_h^j)\right)^{J_j} \left(\|I\|_2 + \|A_i\|_2^{J_i}\right) \tag{C.219}$$

$$\overset{\text{(vi)}}{\leq} 2\,H\,\sqrt{d_{\boldsymbol{c}}\,|\mathcal{D}^t|} \sum_{i=1}^{t} \prod_{j=i+1}^{t} \left(1 - \alpha_{\boldsymbol{c}}^{h,j}\,\lambda_{\min}(\Lambda_h^j)\right)^{J_j} \left(1 + \left(1 - \alpha_{\boldsymbol{c}}^{i}\,\lambda_{\min}(\Lambda_h^i)\right)^{J_i}\right) \tag{C.220}$$

$$\leq 2\,H\,\sqrt{d_{\boldsymbol{c}}\,|\mathcal{D}^t|} \sum_{i=1}^{t} \left(\prod_{j=i+1}^{t} \left(1 - \alpha_{\boldsymbol{c}}^{h,j}\,\lambda_{\min}(\Lambda_h^j)\right)^{J_j} + \prod_{j=i}^{t} \left(1 - \alpha_{\boldsymbol{c}}^{h,j}\,\lambda_{\min}(\Lambda_h^j)\right)^{J_j}\right) \tag{C.221}$$

$$\overset{\text{(vii)}}{\leq} 2\,H\,\sqrt{d_{\boldsymbol{c}}\,|\mathcal{D}^t|} \sum_{i=1}^{t} \left(\prod_{j=i+1}^{t} \left(1 - 1/(2\,\kappa_j)\right)^{J_j} + \prod_{j=i}^{t} \left(1 - 1/(2\,\kappa_j)\right)^{J_j}\right), \tag{C.222}$$

(i) uses the definition of the matrix norm (i.e., $\|A\|_2 \coloneqq \max_x \frac{\|A\,x\|_2}{\|x\|} \implies \|A\,x\|_2 \leq \|A\|_2\,\|x\|_2$). (ii) uses Lemma 18 and sets $\lambda = 1$. (iii) and (v) come from the submultiplicativity of matrix norm. (iv) and (vi) use the fact that $\|A\|_2 \leq \lambda_{\max}(A)$, and (iv) also uses the triangular inequality. (vii) uses the fact that we set $\alpha_{\boldsymbol{c}}^{h,j} = 1/\left(2\,\lambda_{\max}(\Lambda_h^j)\right)$ and denotes that $\kappa_j = \max_{h\in[H]} \lambda_{\max}(\Lambda_h^j)/\lambda_{\min}(\Lambda_h^j)$.

Using Lemma 28. we can set $J_j \geq 2\,\kappa_j \log(1/\sigma)$ where $\sigma = 1/\left(4\,H\,\left(\left|\mathcal{D}^t\right|+1\right)\,\sqrt{d_{\boldsymbol{c}}}\right)$. We can further get that

$$\|\mu_h^{t,J_t}\|_2 \leq 2\,H\,\sqrt{d_{\boldsymbol{c}}\,\left|\mathcal{D}^t\right|}\,\sum_{i=1}^{t}\left(\sigma^{t-i}+\sigma^{t-i+1}\right) \tag{C.223}$$

$$\leq 4\,H\,\sqrt{d_{\boldsymbol{c}}\,\left|\mathcal{D}^t\right|}\,\sum_{i=0}^{\infty}\sigma^{i} \tag{C.224}$$

$$= 4\,H\,\sqrt{d_{\boldsymbol{c}}\,\left|\mathcal{D}^t\right|}\,\left(\frac{1}{1-\sigma}\right) \tag{C.225}$$

$$= \frac{16}{3}\,H\,\sqrt{d_{\boldsymbol{c}}\,\left|\mathcal{D}^t\right|}\,. \tag{C.226}$$

Next, we continue to bound Term (II). Since $\xi_h^{t,J_t} \sim \mathbf{N}(0,\Sigma_h^{t,J_t})$, using Lemma 23, we have that

$$\Pr\left(\left\|\xi_h^{t,J_t}\right\|_2 \leq \sqrt{\frac{1}{\delta}\,\mathrm{Tr}\left(\Sigma_h^{t,J_t}\right)}\right) \geq 1-\delta\,. \tag{C.227}$$

Recall from Lemma 15 that

$$\Sigma_h^{t,J_t} = \sum_{i=1}^{t}\frac{1}{\zeta}A_t^{J_t}\ldots A_{i+1}^{J_{i+1}}\left(I-A_i^{2J_i}\right)\left(\Lambda_h^i\right)^{-1}(I+A_i)^{-1}\,A_{i+1}^{J_{i+1}}\ldots A_t^{J_t}\,. \tag{C.228}$$

Therefore, we can use Lemma 25 and derive that

$$\mathrm{Tr}\left(\Sigma_h^{t,J_t}\right) = \sum_{i=1}^{t}\frac{1}{\zeta}\,\mathrm{Tr}\left(A_t^{J_t}\ldots A_{i+1}^{J_{i+1}}\left(I-A_i^{2J_i}\right)\left(\Lambda_h^i\right)^{-1}(I+A_i)^{-1}\,A_{i+1}^{J_{i+1}}\ldots A_t^{J_t}\right) \tag{C.229}$$

$$\leq \sum_{i=1}^{t}\frac{1}{\zeta}\,\mathrm{Tr}\left(A_t^{J_t}\right)\ldots\mathrm{Tr}\left(A_{i+1}^{J_{i+1}}\right)\mathrm{Tr}\left(I-A_i^{2J_i}\right)\mathrm{Tr}\left(\left(\Lambda_h^i\right)^{-1}\right)\mathrm{Tr}\left((I+A_i)^{-1}\right)\mathrm{Tr}\left(A_{i+1}^{J_{i+1}}\right)\ldots\mathrm{Tr}\left(A_t^{J_t}\right)\,. \tag{C.230}$$

To bound each term, we first have,

$$\mathrm{Tr}\left(A_i^{J_i}\right) \leq \mathrm{Tr}\left(\left(1-\alpha_{\boldsymbol{c}}^i\,\lambda_{\min}(\Lambda_h^i)\right)^{J_i}\,I\right) \tag{C.231}$$

$$\leq d_{\boldsymbol{c}}\,\left(1-\alpha_{\boldsymbol{c}}^i\,\lambda_{\min}(\Lambda_h^i)\right)^{J_i} \tag{C.232}$$

$$\leq d_{\boldsymbol{c}}\,\sigma \leq 1\,, \tag{C.233}$$

where the first inequality follows from the fact that $A_i^{J_i} \prec \left(1-\alpha_{\boldsymbol{c}}^t\lambda_{\min}(\Lambda_h^t)\right)^{J_j}\,I$. Similarly, since we set $0<\alpha_{\boldsymbol{c}}^{h,j}<1/\left(2\,\lambda_{\max}(\Lambda_j)\right)$, we have $A_i^{J_i} \succ \frac{1}{2^{J_i}}I$ and therefore,

$$\mathrm{Tr}\left(I-A_i^{2J_i}\right) \leq \left(1-\frac{1}{2^{2J_i}}\right)\,d_{\boldsymbol{c}} < d_{\boldsymbol{c}}\,. \tag{C.234}$$

Similarly, since we set $0 < \alpha_{\boldsymbol{c}}^{h,j} < 1/\left(2\,\lambda_{\max}(\Lambda_j)\right)$ and thus $I + A_i \succ \frac{3}{2} I$, we have that

$$\operatorname{Tr}\left((I + A_i)^{-1}\right) \leq \frac{2}{3}\, d_{\boldsymbol{c}}\,. \tag{C.235}$$

Additionally, since all eigenvalues of Λ_h^i are greater than or equal to 1,

$$\operatorname{Tr}\left((\Lambda_h^i)^{-1}\right) \leq d_{\boldsymbol{c}} \cdot 1 = d_{\boldsymbol{c}}\,. \tag{C.236}$$

Finally, we have that

$$\operatorname{Tr}\left(\Sigma_h^{t,J_t}\right) \leq \sum_{i=1}^{t} \frac{1}{\zeta} \cdot \frac{2}{3} \cdot d_{\boldsymbol{c}}^3 = \frac{2\, d_{\boldsymbol{c}}^3}{3\,\zeta}\, t\,. \tag{C.237}$$

Therefore, using Lemma 23, we have that

$$\Pr\left(\left\|\xi_h^{t,J_t}\right\|_2 \leq \sqrt{\frac{1}{\delta} \cdot \frac{2\, d_{\boldsymbol{c}}^3}{3\,\zeta}\, T}\right) \geq \Pr\left(\left\|\xi_h^{t,J_t}\right\|_2 \leq \sqrt{\frac{1}{\delta} \operatorname{Tr}\left(\Sigma_h^{t,J_t}\right)}\right) \geq 1 - \delta\,. \tag{C.238}$$

Putting everything together, with probability at least $1 - \delta$, we can obtain that

$$\left\|w_h^{t,m,J_t}\right\|_2 \leq \overline{W}_\delta \coloneqq \frac{16}{3}\, H \sqrt{d_{\boldsymbol{c}}\, |\mathcal{D}^t|} + \sqrt{\frac{2\, d_{\boldsymbol{c}}^3\, t}{3\,\zeta\,\delta}}\,. \tag{C.239}$$

This concludes the proof. □

Proof of Lemma 20

Proof. To start, we decompose the LHS using the triangle inequality,

$$\left|\left\langle \phi(s,a), w_h^{t,m,J_t} - \widehat{w}_h^t \right\rangle\right| \leq \underbrace{\left|\left\langle \phi(s,a), w_h^{t,m,J_t} - \mu_h^{t,J_t} \right\rangle\right|}_{\text{(I)}} + \underbrace{\left|\left\langle \phi(s,a), \mu_h^{t,J_t} - \widehat{w}_h^t \right\rangle\right|}_{\text{(II)}}, \tag{C.240}$$

where μ_h^{t,J_t} is defined in Equation (C.92). To bound Term (I), we first apply Hölder's inequality and obtain that

$$\left|\left\langle \phi(s,a), w_h^{t,m,J_t} - \mu_h^{t,J_t} \right\rangle\right| \leq \left\|\phi(s,a)^\top \left(\Sigma_h^{t,J_t}\right)^{1/2}\right\|_2 \left\|\left(\Sigma_h^{t,J_t}\right)^{-1/2} \left(w_h^{t,m,J_t} - \mu_h^{t,J_t}\right)\right\|_2. \tag{C.241}$$

Since $w_h^{t,m,J_t} \sim \mathbf{N}(\mu_h^{t,J_t}, \Sigma_h^{t,J_t})$, we know that $\left(\Sigma_h^{t,J_t}\right)^{-1/2} \left(w_h^{t,m,J_t} - \mu_h^{t,J_t}\right) \sim \mathbf{N}(0, I_{d_{\boldsymbol{c}} \times d_{\boldsymbol{c}}})$. Therefore,

$$\Pr\left(\left\|\left(\Sigma_h^{t,J_t}\right)^{-1/2} \left(w_h^{t,m,J_t} - \mu_h^{t,J_t}\right)\right\|_2 \geq 2\sqrt{d_{\boldsymbol{c}} \log\left(1/\delta\right)}\right) \leq \delta^2\,. \tag{C.242}$$

Then, we continue to bound $\left\|\phi(s,a)^\top \left(\Sigma_h^{t,J_t}\right)^{1/2}\right\|_2$.

$$\phi(s,a)^\top \Sigma_h^{t,J_t}\, \phi(s,a) \tag{C.243}$$

$$= \frac{1}{\zeta}\sum_{i=1}^{t} \phi(s,a)^\top A_t^{J_t} \dots A_{i+1}^{J_{i+1}} \left(I - A^{2J_i}\right) \left(\Lambda_h^i\right)^{-1} (I+A_i)^{-1} A_{i+1}^{J_{i+1}} \dots A_t^{J_t}\, \phi(s,a) \tag{C.244}$$

$$\overset{\text{(i)}}{=} \frac{1}{\zeta}\sum_{i=1}^{t} \phi(s,a)^\top \mathscr{A}_{i+1} \left(I - A^{2J_i}\right) \left(\Lambda_h^i\right)^{-1} (I+A_i)^{-1} \mathscr{A}_{i+1}^\top \phi(s,a) \tag{C.245}$$

$$\overset{\text{(ii)}}{\leq} \frac{2}{3\zeta}\sum_{i=1}^{t} \phi(s,a)^\top \mathscr{A}_{i+1} \left(\left(\Lambda_h^i\right)^{-1} - A_i^{J_i} \left(\Lambda_h^i\right)^{-1} A_i^{J_i}\right) \mathscr{A}_{i+1}^\top \phi(s,a) \tag{C.246}$$

$$= \frac{2}{3\,\zeta}\left(\sum_{i=1}^{t} \phi(s,a)^\top \mathscr{A}_{i+1} \left(\Lambda_h^i\right)^{-1} \mathscr{A}_{i+1}^\top \phi(s,a) - \sum_{i=1}^{t} \phi(s,a)^\top \mathscr{A}_i \left(\Lambda_h^i\right)^{-1} \mathscr{A}_i^\top \phi(s,a)\right) \tag{C.247}$$

$$\overset{\text{(iii)}}{\leq} \frac{2}{3\,\zeta}\sum_{i=1}^{t} \phi(s,a)^\top \mathscr{A}_{i+1} \left(\Lambda_h^i\right)^{-1} \mathscr{A}_{i+1}^\top \phi(s,a) \tag{C.248}$$

$$= \frac{2}{3\,\zeta}\left(\|\phi(s,a)\|^2_{\left(\Lambda_h^i\right)^{-1}} + \sum_{i=1}^{t-1} \left\|\mathscr{A}_{i+1}^\top \phi(s,a)\right\|^2_{\left(\Lambda_h^i\right)^{-1}}\right) \tag{C.249}$$

$$\leq \frac{2}{3\,\zeta}\,\|\phi(s,a)\|^2_{\left(\Lambda_h^t\right)^{-1}} + \frac{2}{3\,\zeta}\sum_{i=1}^{t-1}\prod_{j=i+1}^{t}\left(1 - \alpha_{\boldsymbol{c}}\,\lambda_{\min}(\Lambda_h^j)\right)^{2\,J_j} \|\phi(s,a)\|^2_{\left(\Lambda_h^i\right)^{-1}}\,. \tag{C.250}$$

For (i), we use the denotation that $\mathscr{A}_{i+1} = A_t^{J_t} \dots A_{i+1}^{J_{i+1}}$. (ii) follows from $I + A_i \succ \frac{3}{2} I$ since we set $\alpha_{\boldsymbol{c}}^{h,j} = 1/\left(2\,\lambda_{\max}(\Lambda_h^j)\right)$. (iii) follows from the fact that $\sum_{i=1}^{t} \phi(s,a)^\top \mathscr{A}_i \left(\Lambda_h^i\right)^{-1} \mathscr{A}_i^\top \phi(s,a) > 0$. Therefore,

$$\left\|\phi(s,a)^\top \left(\Sigma_h^{t,J_t}\right)^{1/2}\right\|_2 = \sqrt{\phi(s,a)^\top \Sigma_h^{t,J_t}\, \phi(s,a)} \tag{C.251}$$

$$\overset{\text{(iv)}}{\leq} \sqrt{\frac{2}{3\,\zeta}}\,\|\phi(s,a)\|_{\left(\Lambda_h^t\right)^{-1}} + \sqrt{\frac{2}{3\,\zeta}}\sum_{i=1}^{t-1}\prod_{j=i+1}^{t}\left(1 - \alpha_{\boldsymbol{c}}\,\lambda_{\min}(\Lambda_h^j)\right)^{J_j} \|\phi(s,a)\|_{\left(\Lambda_h^i\right)^{-1}} \tag{C.252}$$

$$\overset{\text{(v)}}{\leq} \sqrt{\frac{2}{3\,\zeta}}\,\|\phi(s,a)\|_{\left(\Lambda_h^t\right)^{-1}} + \sqrt{\frac{2}{3\,\zeta}}\sum_{i=1}^{t-1} \sigma^{t-i}\, \|\phi(s,a)\|_{\left(\Lambda_h^i\right)^{-1}} \tag{C.253}$$

$$\overset{\text{(vi)}}{\leq} \sqrt{\frac{2}{3\,\zeta}}\,\|\phi(s,a)\|_{\left(\Lambda_h^t\right)^{-1}} + \sqrt{\frac{2\,\left(|\mathcal{D}^t|+1\right)}{3\,\zeta}}\left(\sum_{i=1}^{t-1} \sigma^{t-i}\right) \|\phi(s,a)\|_{\left(\Lambda_h^t\right)^{-1}} \tag{C.254}$$

$$\overset{\text{(vii)}}{\leq} \sqrt{\frac{2}{3\,\zeta}}\,\|\phi(s,a)\|_{\left(\Lambda_h^t\right)^{-1}} + \sqrt{\frac{2\,\left(|\mathcal{D}^t|+1\right)}{3\,\zeta}}\left(\frac{\sigma}{1-\sigma}\right) \|\phi(s,a)\|_{\left(\Lambda_h^t\right)^{-1}} \tag{C.255}$$

$$\leq \sqrt{\frac{2}{3\,\zeta}}\,\|\phi(s,a)\|_{\left(\Lambda_h^t\right)^{-1}} + \frac{1}{4}\sqrt{\frac{2}{3\,\zeta}}\left(\frac{1}{1-\sigma}\right) \|\phi(s,a)\|_{\left(\Lambda_h^t\right)^{-1}} \tag{C.256}$$

$$\leq \left(\sqrt{\frac{2}{3\,\zeta}} + \frac{1}{3}\sqrt{\frac{2}{3\,\zeta}}\right) \|\phi(s,a)\|_{(\Lambda_h^t)^{-1}} \tag{C.257}$$

$$\leq \frac{4}{3}\sqrt{\frac{2}{3\,\zeta}}\, \|\phi(s,a)\|_{(\Lambda_h^t)^{-1}}\,. \tag{C.258}$$

(iv) follows from the fact that $\sqrt{a+b} \leq a+b$ for all $a, b > 0$. (v) uses Lemma 28 by setting $J_j \geq \kappa_j \log(1/\sigma)$ where $\sigma = 1/\left(4\,H\,\left(\left|\mathcal{D}^t\right|+1\right)\,\sqrt{d_{\boldsymbol{c}}}\right)$. (vi) follows from $\|\phi(s,a)\|_{(\Lambda_h^i)^{-1}} \leq \|\phi(s,a)\|_2 \leq \sqrt{|\mathcal{D}^t|+1}\,\|\phi(s,a)\|_{(\Lambda_h^t)^{-1}}$. (vii) follows from $\sum_{i=1}^t \sigma^{t-i} \leq \sum_{i=1}^\infty \sigma^i \leq \sigma/(1-\sigma)$. Therefore, we have

$$\Pr\left(\left|\left\langle \phi(s,a), w_h^{t,m,J_t} - \mu_h^{t,J_t}\right\rangle\right| \geq \frac{8}{3}\sqrt{\frac{2\,d_{\boldsymbol{c}}\,\log(1/\delta)}{3\,\zeta}}\,\|\phi(s,a)\|_{(\Lambda_h^t)^{-1}}\right) \tag{C.259}$$

$$\leq \Pr\left(\left\|\phi(s,a)^\top\left(\Sigma_h^{t,J_t}\right)^{1/2}\right\|_2 \left\|\left(\Sigma_h^{t,J_t}\right)^{-1/2}\left(w_h^{t,m,J_t} - \mu_h^{t,J_t}\right)\right\|_2 \geq 2\sqrt{d_{\boldsymbol{c}}\,\log(1/\delta)}\,\left\|\phi(s,a)^\top\left(\Sigma_h^{t,J_t}\right)^{1/2}\right\|_2\right) \tag{C.260}$$

$$= \Pr\left(\left\|\left(\Sigma_h^{t,J_t}\right)^{-1/2}\left(w_h^{t,m,J_t} - \mu_h^{t,J_t}\right)\right\|_2 \geq 2\sqrt{d_{\boldsymbol{c}}\,\log(1/\delta)}\right) = \delta^2 \leq \delta\,. \tag{C.261}$$

This implies that

$$\Pr\left(\left|\left\langle \phi(s,a), w_h^{t,m,J_t} - \mu_h^{t,J_t}\right\rangle\right| \leq \frac{8}{3}\sqrt{\frac{2\,d_{\boldsymbol{c}}\,\log(1/\delta)}{3\,\zeta}}\,\|\phi(s,a)\|_{(\Lambda_h^t)^{-1}}\right) \geq 1-\delta\,. \tag{C.262}$$

Putting everything together, with probability at least $1-\delta$,

$$\left|\left\langle \phi(s,a), w_h^{t,m,J_t} - \widehat{w}_h^t\right\rangle\right| \leq \left|\left\langle \phi(s,a), w_h^{t,m,J_t} - \mu_h^{t,J_t}\right\rangle\right| + \left|\left\langle \phi(s,a), \mu_h^{t,J_t} - \widehat{w}_h^t\right\rangle\right| \tag{C.263}$$

$$\leq \left(\frac{8}{3}\sqrt{\frac{2\,d_{\boldsymbol{c}}\,\log(1/\delta)}{3\,\zeta}} + \frac{4}{3}\right)\|\phi(s,a)\|_{(\Lambda_h^t)^{-1}}\,. \tag{C.264}$$

□

Proof of Lemma 21

Proof. Recall that $\mathbb{P}_h V(s,a) \coloneqq \mathbb{E}_{s'\sim\mathbb{P}_h(\cdot|s,a)} V(s')$ and $\mathbb{P}_h(\cdot \mid s,a) = \langle \phi(s,a), \psi_h(\cdot)\rangle$ due to the linear MDP assumption (Definition 1). We also denote that $\widehat{\Psi}_h^t \coloneqq \left\langle \psi_h, \widehat{V}_{h+1}^t\right\rangle_{\mathcal{S}}$ and thus $\mathbb{P}_h \widehat{V}_{h+1}^t(s,a) = \left\langle \phi(s,a), \widehat{\Psi}_h^t\right\rangle$. Then, we have that

$$\mathbb{P}_h \widehat{V}_{h+1}^t(s,a) = \left\langle \phi(s,a), \widehat{\Psi}_h^t\right\rangle \tag{C.265}$$

$$= \phi(s,a)^\top\left(\Lambda_h^t\right)^{-1}\Lambda_h^t\,\widehat{\Psi}_h^t \tag{C.266}$$

$$= \phi(s,a)^\top\left(\Lambda_h^t\right)^{-1}\left(\sum_{(s,a,s')\in\mathcal{D}_h^t}\phi(s,a)\,\phi(s,a)^\top + \lambda I\right)\widehat{\Psi}_h^t \tag{C.267}$$

$$= \phi(s,a)^\top \left(\Lambda_h^t\right)^{-1} \left(\sum_{(s,a,s') \in \mathcal{D}_h^t} \phi(s,a) (\mathbb{P}_h \widehat{V}_{h+1}^t)(s,a) + \lambda \, \widehat{\Psi}_h^t \right) . \tag{C.268}$$

This further implies that

$$\left\langle \phi(s,a), \widehat{w}_h^t \right\rangle - r_h(s,a) - \mathbb{P}_h \widehat{V}_{h+1}^t(s,a) \tag{C.269}$$

$$= \phi(s,a)^\top \left(\Lambda_h^t\right)^{-1} \sum_{(s,a,s') \in \mathcal{D}_h^t} \left[r_h(s,a) + \widehat{V}_{h+1}^t(s') \right] \cdot \phi(s,a) - r_h(s,a) \tag{C.270}$$

$$- \phi(s,a)^\top \left(\Lambda_h^t\right)^{-1} \left(\sum_{(s,a,s') \in \mathcal{D}_h^t} \phi(s,a) (\mathbb{P}_h \widehat{V}_{h+1}^t)(s,a) + \lambda \, \widehat{\Psi}_h^t \right) \tag{C.271}$$

$$= \underbrace{\phi(s,a)^\top (\Lambda_h^t)^{-1} \left(\sum_{(s,a,s') \in \mathcal{D}_h^t} \phi(s,a) \left[\widehat{V}_{h+1}^t(s') - \mathbb{P}_h \widehat{V}_{h+1}^t(s,a) \right] \right)}_{\text{(I)}} \tag{C.272}$$

$$+ \underbrace{\phi(s,a)^\top (\Lambda_h^t)^{-1} \left(\sum_{(s,a,s') \in \mathcal{D}_h^t} r_h(s,a) \, \phi(s,a) \right) - r_h(s,a)}_{\text{(II)}} \underbrace{- \lambda \, \phi(s,a)^\top (\Lambda_h^t)^{-1} \, \widehat{\Psi}_h^t}_{\text{(III)}} . \tag{C.273}$$

We first start by bounding Term (I). With probability at least $1 - \delta$, it holds that

$$\phi(s,a)^\top (\Lambda_h^t)^{-1} \left(\sum_{(s,a,s') \in \mathcal{D}_h^t} \phi(s,a) \left[\widehat{V}_{h+1}^t(s') - \mathbb{P}_h \widehat{V}_{h+1}^t(s,a) \right] \right) \tag{C.274}$$

$$\overset{\text{(i)}}{\leq} \left\| \sum_{(s,a,s') \in \mathcal{D}_h^t} \phi(s,a) \left[\widehat{V}_{h+1}^t(s') - \mathbb{P}_h \widehat{V}_{h+1}^t(s,a) \right] \right\|_{(\Lambda_h^t)^{-1}} \|\phi(s,a)\|_{(\Lambda_h^t)^{-1}} \tag{C.275}$$

$$\overset{\text{(ii)}}{\leq} C_\delta \, H \, \sqrt{d_{\boldsymbol{c}}} \, \|\phi(s,a)\|_{(\Lambda_h^t)^{-1}} , \tag{C.276}$$

where (i) follows from the Cauchy-Schwarz inequality, and (ii) follows from the good event defined in Lemma 13.

Next, we continue to bound Term (II). We observe that

$$\phi(s,a)^\top (\Lambda_h^t)^{-1} \left(\sum_{(s,a,s') \in \mathcal{D}_h^t} r_h(s,a) \, \phi(s,a) \right) - r_h(s,a) \tag{C.277}$$

$$\overset{\text{(iii)}}{=} \phi(s,a)^\top (\Lambda_h^t)^{-1} \left(\sum_{(s,a,s') \in \mathcal{D}_h^t} r_h(s,a) \, \phi(s,a) \right) - \phi(s,a)^\top \theta_h \tag{C.278}$$

$$= \phi(s,a)^\top (\Lambda_h^t)^{-1} \left(\sum_{(s,a,s') \in \mathcal{D}_h^t} r_h(s,a)\, \phi(s,a) - \Lambda_h^t \theta_h \right) \tag{C.279}$$

$$= \phi(s,a)^\top (\Lambda_h^t)^{-1} \left(\sum_{(s,a,s') \in \mathcal{D}_h^t} r_h(s,a)\, \phi(s,a) - \sum_{(s,a,s') \in \mathcal{D}_h^t} \phi(s,a)\, \phi(s,a)^\top \theta_h - \lambda\, \theta_h \right) \tag{C.280}$$

$$\overset{\text{(iv)}}{=} \phi(s,a)^\top (\Lambda_h^t)^{-1} \left(\sum_{(s,a,s') \in \mathcal{D}_h^t} r_h(s,a)\, \phi(s,a) - \sum_{(s,a,s') \in \mathcal{D}_h^t} \phi(s,a)\, r_h(s,a) - \lambda\, \theta_h \right) \tag{C.281}$$

$$= -\lambda\, \phi(s,a)^\top\, (\Lambda_h^t)^{-1}\, \theta_h \tag{C.282}$$

$$\overset{\text{(v)}}{\leq} \lambda\, \|\phi(s,a)\|_{(\Lambda_h^t)^{-1}}\, \|\theta_h\|_{(\Lambda_h^t)^{-1}} \tag{C.283}$$

$$\overset{\text{(vi)}}{\leq} \sqrt{\lambda\, d_{\boldsymbol{c}}}\, \|\phi(s,a)\|_{(\Lambda_h^t)^{-1}}\ . \tag{C.284}$$

(iii) and (iv) follow from the definition $r_h(s,a) = \langle \phi(s,a), \theta_h \rangle$. (v) applies the Cauchy-Schwarz inequality. (vi) follows from $\|\phi(s,a)\|_{(\Lambda_h^t)^{-1}} \leq \sqrt{1/\lambda}\, \|\phi(s,a)\|_2$ and $\|\theta_h\|_2 \leq \sqrt{d_{\boldsymbol{c}}}$ (Definition 1).

Lastly, we derive the bound for Term (III).

$$\lambda\, \phi(s,a)^\top\, (\Lambda_h^t)^{-1}\, \widehat{\Psi}_h^t \overset{\text{(vii)}}{\leq} \lambda\, \|\phi(s,a)\|_{(\Lambda_h^t)^{-1}}\, \left\|\widehat{\Psi}_h^t\right\|_{(\Lambda_h^t)^{-1}} \tag{C.285}$$

$$\overset{\text{(viii)}}{\leq} \sqrt{\lambda}\, \|\phi(s,a)\|_{(\Lambda_h^t)^{-1}}\, \left\|\widehat{\Psi}_h^t\right\|_2 \tag{C.286}$$

$$\leq \sqrt{\lambda}\, \|\phi(s,a)\|_{(\Lambda_h^t)^{-1}}\, \left\|\left\langle \psi_h, \widehat{V}_{h+1}^t \right\rangle_{\mathcal{S}}\right\|_2 \tag{C.287}$$

$$= H\, \sqrt{\lambda}\, \|\phi(s,a)\|_{(\Lambda_h^t)^{-1}}\, \left\| \int_{s \in \mathcal{S}} \psi_h(s) \left(\widehat{V}_{h+1}^t(s)/H \right) \mathrm{d}\, s \right\|_2 \tag{C.288}$$

$$\overset{\text{(xiv)}}{\leq} H\, \sqrt{\lambda\, d_{\boldsymbol{c}}}\, \|\phi(s,a)\|_{(\Lambda_h^t)^{-1}}\ . \tag{C.289}$$

(vii) applies the Cauchy-Schwarz inequality. (viii) follows from $\left\|\widehat{\Psi}_h^t\right\|_{(\Lambda_h^t)^{-1}} \leq \sqrt{\lambda}\, \left\|\widehat{\Psi}_h^t\right\|_2$. (xiv) comes from the assumption that $\left\| \int_{s \in \mathcal{S}} \psi_h(s) \left(\widehat{V}_{h+1}^t(s)/H \right) \mathrm{d}\, s \right\|_2 \leq \sqrt{d_{\boldsymbol{c}}}$ (Definition 1).

Putting everything together and setting $\lambda = 1$, we have with probability at least $1 - \delta$,

$$\left| \left\langle \phi(s,a), \widehat{w}_h^t \right\rangle - r_h(s,a) - \mathbb{P}_h \widehat{V}_{h+1}^t(s,a) \right| \tag{C.290}$$

$$\leq \left(C_\delta\, H\, \sqrt{d_{\boldsymbol{c}}} + \sqrt{\lambda\, d_{\boldsymbol{c}}} + H\, \sqrt{\lambda\, d_{\boldsymbol{c}}} \right)\, \|\phi(s,a)\|_{(\Lambda_h^t)^{-1}} \tag{C.291}$$

$$= 3\, C_\delta\, H\, \sqrt{d_{\boldsymbol{c}}}\, \|\phi(s,a)\|_{(\Lambda_h^t)^{-1}}\ . \tag{C.292}$$

This concludes the proof. □

C.4.3 Technical Tools

Lemma 22 (Jin et al. 2020, Lemma D.1)**.** *Let $\Lambda = \lambda I + \sum_{i=1}^{t} \phi_i \phi_i^\top$, where $\phi_i \in \mathbb{R}^d$ and $\lambda > 0$. Then,*

$$\sum_{i=1}^{t} \phi_i^\top (\Lambda)^{-1} \phi_i \leq d\,. \tag{C.293}$$

Lemma 23 (Ishfaq et al. 2024a, Lemma E.1)**.** *Given a multivariate normal distribution $X \sim \mathbf{N}(0, \Sigma_{d\times d})$, for any $\delta \in (0, 1]$, it hold that*

$$\Pr\left(\|X\|_2 \leq \sqrt{\frac{1}{\delta} \operatorname{Tr}(\Sigma)} \right) \geq 1 - \delta\,. \tag{C.294}$$

Lemma 24 (Abramowitz and Stegun 1948)**.** *Suppose X is a Gaussian random variable $X \sim \mathbf{N}(\mu, \sigma^2)$, where $\sigma > 0$. For $z \in [0, 1]$, it holds that*

$$\Pr(X > \mu + z\,\sigma) \geq \frac{e^{-z^2/2}}{\sqrt{8\pi}} \quad \textit{and} \quad \Pr(X < \mu - z\,\sigma) \geq \frac{e^{-z^2/2}}{\sqrt{8\pi}}\,. \tag{C.295}$$

Additionally, for any $z \geq 1$,

$$\frac{e^{-z^2/2}}{2\,z\,\sqrt{\pi}} \leq \Pr(|X - \mu| > z\,\sigma) \leq \frac{e^{-z^2/2}}{z\,\sqrt{\pi}}\,. \tag{C.296}$$

Lemma 25. *If A and B are positive semi-definite square matrices of the same size, then*

$$[\operatorname{Tr}(AB)]^2 \leq \operatorname{Tr}(A^2)\operatorname{Tr}(B^2) \leq [\operatorname{Tr}(A)]^2 [\operatorname{Tr}(B)]^2\,. \tag{C.297}$$

Lemma 26. *Given two symmetric positive semi-definite square matrices A and B such that $A \succeq B$, it holds that $\|A\|_2 \geq \|B\|_2$.*

Proof. Note that $A - B$ is also positive semi-definite. Then, we have that

$$\|B\|_2 = \sup_{\|x\|=1} x^\top B x \leq \sup_{\|x\|=1} \left(x^\top B x + x^\top (A - B) x \right) = \sup_{\|x\|=1} x^\top A x = \|A\|_2\,. \tag{C.298}$$

□

Lemma 27. *Let $A \in \mathbb{R}^{d\times d}$ be a positive definite matrix where its largest eigenvalue $\lambda_{\max}(A) \leq \lambda$. Given that $v_1, \ldots, v_n$ are n vectors in $\mathbb{R}^d$, it holds that*

$$\left\| A \sum_{i=1}^{n} v_i \right\| \leq \sqrt{\lambda\, n \sum_{i=1}^{n} \|v_i\|_A^2}\,. \tag{C.299}$$

Lemma 28. *Let Λ be a positive definite matrix and $\kappa = \frac{\lambda_{\max}(\Lambda)}{\lambda_{\min}(\Lambda)}$ be the condition number of Λ. If $\Lambda \succ I$ and $J \geq 2\,\kappa \log(1/\sigma)$, then, for any $\sigma > 0$,*

$$\left(1 - 1/\left(2\,\kappa\right)\right)^J < \sigma\,. \tag{C.300}$$

Proof. The statement is equivalent to proving that

$$J \geq \frac{\log\left(1/\sigma\right)}{\log\left(\frac{1}{1-1/\left(2\,\kappa\right)}\right)}\,. \tag{C.301}$$

Since $\kappa \geq 1$ and for any $x \in (0,1)$, $e^{-x} > 1-x$, we have that

$$e^{-1/\left(2\,\kappa\right)} > 1 - 1/\left(2\,\kappa\right) \implies \log\left(\frac{1}{1-1/\left(2\,\kappa\right)}\right) \geq \frac{1}{2\,\kappa}\,. \tag{C.302}$$

Therefore, we have that

$$J \geq 2\,\kappa\,\log\left(1/\sigma\right) \geq \frac{\log(1/\sigma)}{\log\left(\frac{1}{1-1/\left(2\,\kappa\right)}\right)}\,. \tag{C.303}$$

This concludes the proof. □

C.5 Sample Complexity in the On-Policy Setting

C.5.1 Proof of Good Event

Lemma 29. *Consider Algorithm 2 in the on-policy setting with $\lambda = 1$. Then, for any $\delta \in (0,1)$, with probability at least $1-\delta$, it holds that*

$$\left\| \sum_{(s,a,s') \in \mathcal{D}_h^t} \phi(s,a) \left[\widehat{V}_{h+1}^t(s') - \mathbb{P}_h \widehat{V}_{h+1}^t(s,a) \right] \right\|_{(\Lambda_h^t)^{-1}} \leq C_\delta^{\mathrm{on}}\, H \sqrt{d_{\boldsymbol{c}}}\,, \tag{C.304}$$

where $C_\delta^{\mathrm{on}} = \log(N/\delta)$.

Proof of Lemma 29. Recall that $\mathbb{P}_h \widehat{V}_{h+1}^t(s,a) = \mathbb{E}_{s' \sim \mathbb{P}_h}\left[\widehat{V}_{h+1}^t(s')\right]$. Thus, $\mathbb{E}[\widehat{V}_{h+1}^t(s') - \mathbb{P}_h \widehat{V}_{h+1}^t(s,a)] = 0$. Also, $\left|\widehat{V}_{h+1}^t(s') - \mathbb{P}_h \widehat{V}_{h+1}^t(s,a)\right| \leq H$. Therefore, $\widehat{V}_{h+1}^t(s') - \mathbb{P}_h \widehat{V}_{h+1}^t(s,a)$ is zero-mean and H-sub Gaussian. Given that, we can invoke Lemma 31.

$$\left\| \sum_{(s,a,s') \in \mathcal{D}_h^t} \phi(s,a) \left[\widehat{V}_{h+1}^t(s') - \mathbb{P}_h \widehat{V}_{h+1}^t(s,a) \right] \right\|_{(\Lambda_h^t)^{-1}} \tag{C.305}$$

$$\leq \sqrt{2}\, H \sqrt{\log\left[\frac{\det(\Lambda_h^t)^{1/2}\det(\Lambda_h^0)^{-1/2}}{\delta}\right]} \tag{C.306}$$

$$= \sqrt{2}\, H \sqrt{\log\left[\left(\frac{N+\lambda}{\lambda}\right)^{d/2}\right] - \log(\delta)} \tag{C.307}$$

$$= \sqrt{2}\, H \sqrt{\frac{d_{\boldsymbol{c}}}{2}\log(N/\delta)} \tag{C.308}$$

$$= H \sqrt{d_{\boldsymbol{c}}\,\log(N/\delta)}\,, \tag{C.309}$$

where the first equality follows from Lemma 32, and the second equality holds by setting $\lambda = 1$. This concludes the proof. □

C.5.2 Proof of Theorem 7

Using Lemma 29, we can instantiate Lemma 14 in the on-policy setting with

$$\Gamma^{\mathrm{on}}_{\mathtt{LMC}} = C^{\mathrm{on}}_{\delta}\, H\, \sqrt{d_{\boldsymbol{c}}} + \frac{4}{3}\sqrt{\frac{2\, d_{\boldsymbol{c}} \log\left(1/\delta\right)}{3\,\zeta}} + \frac{4}{3} \tag{C.310}$$

$$= H\sqrt{d_{\boldsymbol{c}}\, \log(N/\delta)} + \frac{4}{3}\sqrt{\frac{2\, d_{\boldsymbol{c}} \log\left(1/\delta\right)}{3\,\zeta}} + \frac{4}{3}\,. \tag{C.311}$$

Proof of Theorem 7. The optimal gap for the mixture policy can be written as

$$\mathbb{E}\left[V_1^{\star}(s_1) - V_1^{\overline{\pi}^T}(s_1)\right] = \frac{1}{T}\sum_{t=1}^{T}\left(V_1^{\star}(s_1) - V_1^{\pi^t}(s_1)\right)\,. \tag{C.312}$$

Then, to decompose the above summation, we have that

$$\sum_{t=1}^{T}\left(V_1^{\star}(s_1) - V_1^{\pi^t}(s_1)\right) = \sum_{t=1}^{T}\left(V_1^{\star}(s_1) - \widehat{V}_1^{t}(s_1)\right) + \sum_{t=1}^{T}\left(\widehat{V}_1^{t}(s_1) - V_1^{\pi^t}(s_1)\right)\,. \tag{C.313}$$

We can further decompose the first term by invoking Lemma 30 with $\pi = \pi^{\star}$ and obtain that

$$V_1^{\star}(s_1) - \widehat{V}_1^{t}(s_1) = \sum_{h=1}^{H}\mathbb{E}_{\pi^{\star}}\left[\left\langle \pi_h^{\star}(\cdot \mid s) - \pi_h^{t}(\cdot \mid s), \widehat{Q}_h^{t}(s,\cdot)\right\rangle\right] + \sum_{h=1}^{H}\mathbb{E}_{\pi^{\star}}\left[r_h(s,a) + \mathbb{P}_h\widehat{V}_{h+1}^{t}(s,a) - \widehat{Q}_h^{t}(s,a)\right]\,. \tag{C.314}$$

Similarly, we can decompose the second term by invoking Lemma 30 with $\pi = \pi^t$ and get that

$$\widehat{V}_1^{t}(s_1) - V_1^{\pi^t}(s_1) = \sum_{h=1}^{H}\mathbb{E}_{\pi^{t}}\left[\left\langle \pi_h^{t}(\cdot \mid s) - \pi_h^{t}(\cdot \mid s), \widehat{Q}_h^{t}(s,\cdot)\right\rangle\right] - \sum_{h=1}^{H}\mathbb{E}_{\pi^{t}}\left[r_h(s,a) + \mathbb{P}_h\widehat{V}_{h+1}^{t}(s,a) - \widehat{Q}_h^{t}(s,a)\right] \tag{C.315}$$

$$= -\sum_{h=1}^{H}\mathbb{E}_{\pi^{t}}\left[r_h(s,a) + \mathbb{P}_h\widehat{V}_{h+1}^{t}(s,a) - \widehat{Q}_h^{t}(s,a)\right]\,. \tag{C.316}$$

Therefore, using the definition of the model prediction error ι in Definition 2, we have

$$\sum_{t=1}^{T}\left(V_1^{\star}(s_1) - V_1^{\pi^t}(s_1)\right) \tag{C.317}$$

$$= \underbrace{\sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}_{\pi^{\star}}\left[\left\langle \pi_h^{\star}(\cdot \mid s) - \pi_h^{t}(\cdot \mid s), \widehat{Q}_h^{t}(s,\cdot)\right\rangle\right]}_{\text{(I) policy optimization (actor) error}} + \underbrace{\sum_{t=1}^{T}\sum_{h=1}^{H}\left(\mathbb{E}_{\pi^{\star}}[\iota_h^{t}(s,a)] - \mathbb{E}_{\pi^{t}}[\iota_h^{t}(s,a)]\right)}_{\text{(II) policy evaluation (critic) error}}\,. \tag{C.318}$$

Policy optimization error. We first start by bounding Term (I), the policy optimization (actor) error.

$$\text{Term (I)} = \sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}_{s\sim\pi^\star}\left[\left\langle \pi_h^\star(\cdot \mid s) - \pi_h^t(\cdot \mid s), \widehat{Q}_h^t(s,\cdot)\right\rangle\right] \tag{C.319}$$

$$= \sum_{h=1}^{H} \mathbb{E}_{s\sim\pi^\star}\left(\sum_{t=1}^{T}\left\langle \pi_h^\star(\cdot \mid s) - \pi_h^t(\cdot \mid s), \widehat{Q}_h^t(s,\cdot)\right\rangle\right) \tag{C.320}$$

$$\leq H \max_{(h,s)\in[H]\times\mathcal{S}}\left(\sum_{t=1}^{T}\left\langle \pi_h^\star(\cdot \mid s) - \pi_h^t(\cdot \mid s), \widehat{Q}_h^t(s,\cdot)\right\rangle\right) \tag{C.321}$$

$$\overset{\text{(i)}}{\leq} H\left(\frac{\log|\mathcal{A}| + \sum_{t=1}^{T}\left\|\epsilon_h^t(\cdot)\right\|_\infty}{\eta} + \frac{\eta\, H^2\, T}{2}\right) \tag{C.322}$$

$$\overset{\text{(ii)}}{\leq} H^2\sqrt{\left(\log|\mathcal{A}| + \bar{\epsilon}\, T\right)/2}\,\sqrt{T} \tag{C.323}$$

$$\overset{\text{(iii)}}{\leq} \mathcal{O}\left(H^2\sqrt{\log|\mathcal{A}|}\,\sqrt{T} + H^2\sqrt{\bar{\epsilon}}\, T\right). \tag{C.324}$$

(i) follows from Theorem 6 with $u = \pi_h^*(\cdot \mid s)$. (ii) is obtained by setting $\eta = \frac{\sqrt{2\left(\log|\mathcal{A}| + \bar{\epsilon}\, T\right)}}{H\sqrt{T}}$. (iii) is based on that for all $a, b \geq 0$, $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$.

Policy evaluation error. Then, we continue to bound Term (II), the policy evaluation (critic) error.

$$\text{Term (II)} = \sum_{t=1}^{T}\sum_{h=1}^{H}\left(\mathbb{E}_{\pi^\star}[\iota_h^t(s,a)] - \mathbb{E}_{\pi^t}[\iota_h^t(s,a)]\right) \tag{C.325}$$

$$\overset{\text{(iv)}}{\leq} -\sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}_{\pi^t}[\iota_h^t(s,a)] \tag{C.326}$$

$$\overset{\text{(v)}}{\leq} \Gamma_{\mathtt{LMC}}^{\text{on}}\sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}_{\pi^t}\left[\left\|\phi(s,a)\right\|_{(\Lambda_h^t)^{-1}}\right] \tag{C.327}$$

$$\leq \Gamma_{\mathtt{LMC}}^{\text{on}}\, T \max_{t\in[T]}\sum_{h=1}^{H}\mathbb{E}_{\pi^t}\left[\left\|\phi(s,a)\right\|_{(\Lambda_h^t)^{-1}}\right]. \tag{C.328}$$

(iv) and (v) both follow from Lemma 14, where (iv) is based on the optimism guarantee (RHS of Equation (C.89)), while (v) is based on the error bound (LHS of Equation (C.89)).

Bounding the sum of bonuses. Since $\Gamma_{\mathtt{LMC}}^{\text{on}}$ is bounded, it suffices to bound $\mathbb{E}_{\pi^t}\left[\left\|\phi(s,a)\right\|_{(\Lambda_h^t)^{-1}}\right]$. Note that $\Lambda_h^t = \sum_{(s,a,s')\in\mathcal{D}_h^t}\phi(s,a)\,\phi(s,a)^\top + \lambda\, I$, and $\mathcal{D}_h^t$ only depends on π^t in the on-policy setting. (This is not true for the off-policy setting since Λ_h^t would depend on $\{\pi^1, \ldots, \pi^t\}$.) We then index each data point in $\mathcal{D}_h^t$ as $\left\{(s_h^i, a_h^i, s_{h+1}^i)\right\}_{i\in[N]}$. Let $\Lambda_h^{t,i} =$

$\left(\sum_{j=1}^{i} \phi(s_h^j, a_h^j)\, \phi(s_h^j, a_h^j)^\top + \lambda\, I\right)$. Then, we have that

$$\sum_{h=1}^{H} \mathbb{E}_{\pi^t}\left[\|\phi(s,a)\|_{(\Lambda_h^t)^{-1}}\right] \tag{C.329}$$

$$\overset{\text{(vi)}}{\leq} \frac{1}{N}\sum_{i=1}^{N}\sum_{h=1}^{H} \mathbb{E}_{\pi^t}\left[\|\phi(s,a)\|_{\left(\Lambda_h^{t,i}\right)^{-1}}\right] \tag{C.330}$$

$$= \frac{1}{N}\sum_{i=1}^{N}\sum_{h=1}^{H} \left\|\phi(s_h^i, a_h^i)\right\|_{\left(\Lambda_h^{t,i}\right)^{-1}} + \frac{1}{N}\sum_{i=1}^{N}\sum_{h=1}^{H} \underbrace{\mathbb{E}_{\substack{s_h\sim\mathbb{P}(\cdot|s_{h-1}^t a_{h-1}^t)\\ a_h\sim\pi_h^t(\cdot|s_h)}}\left[\|\phi(s,a)\|_{\left(\Lambda_h^{t,i}\right)^{-1}}\right] - \left\|\phi(s_h^i, a_h^i)\right\|_{\left(\Lambda_h^{t,i}\right)^{-1}}}_{\coloneqq \mathcal{M}_{i,h}^{\text{on}}}, \tag{C.331}$$

where (vi) follows from the fact that $\Lambda_h^{t,i} \preceq \Lambda_h^t$.

Applying the elliptical potential lemma. For the first term of Equation (C.331), we have that

$$\frac{1}{N}\sum_{i=1}^{N}\sum_{h=1}^{H} \left\|\phi(s_h^i, a_h^i)\right\|_{\left(\Lambda_h^{t,i}\right)^{-1}} \tag{C.332}$$

$$= \frac{1}{N}\sum_{h=1}^{H}\sum_{i=1}^{N} \left\|\phi(s_h^i, a_h^i)\right\|_{\left(\Lambda_h^{t,i}\right)^{-1}} \tag{C.333}$$

$$\overset{\text{(vii)}}{\leq} \frac{1}{N}\sum_{h=1}^{H} \sqrt{N}\left(\sum_{i=1}^{N} \left\|\phi(s_h^i, a_h^i)\right\|^2_{\left(\Lambda_h^{t,i}\right)^{-1}}\right)^{1/2} \tag{C.334}$$

$$\overset{\text{(viii)}}{\leq} \mathcal{O}\left(\sqrt{\frac{d_{\boldsymbol{c}}\, H^2 \log(N/\delta)}{N}}\right). \tag{C.335}$$

(vii) applies the Cauchy-Schwarz inequality, and (viii) follows the elliptical argument from Lemma 33.

A martingale difference sequence. For the second term of Equation (C.331), since for a fixed $i \in [N]$, $\left\{\mathcal{M}_{i,h}^{\text{on}}\right\}_{h\in[H]}$ forms a martingale sequence adapted to the filtration,

$$\mathcal{F}_{i,h}^{\text{on}} = \left\{(s_\tau^i, a_\tau^i)\right\}_{\tau\in[h-1]}, \tag{C.336}$$

such that $\mathbb{E}\left[\mathcal{M}_{i,h}^{\text{on}} \mid \mathcal{F}_{i,h}^{\text{on}}\right] = 0$, where the expectation is with respect to the randomness in the policy and the environment at step h. Since $|\mathcal{M}_{i,h}^{\text{on}}| \leq 1$, we can apply the Azuma–Hoeffding inequality and obtain that

$$\Pr\left(\sum_{i=1}^{N}\sum_{h=1}^{H} \mathcal{M}_{i,h}^{\text{on}} \geq m\right) \geq \exp\left(\frac{-m^2}{2\, H\, N}\right). \tag{C.337}$$

Setting $m = \sqrt{2\,H\,N\,\log(1/\delta)}$ and using a union bound over $i \in [N]$, with probability at least $1-\delta$, it holds that

$$\frac{1}{N}\sum_{i=1}^{N}\sum_{h=1}^{H}\mathcal{M}_{i,h}^{\text{on}} \leq \sqrt{\frac{2\,H\log(1/\delta)}{N}} \leq \mathcal{O}\left(\sqrt{\frac{H\,\log(1/\delta)}{N}}\right). \tag{C.338}$$

Putting everything together. Therefore, we have that

$$\sum_{h=1}^{H}\mathbb{E}_{\pi^t}\left[\|\phi(s,a)\|_{(\Lambda_h^t)^{-1}}\right] = \frac{1}{N}\sum_{i=1}^{N}\sum_{h=1}^{H}\left\|\phi(s_h^i,a_h^i)\right\|_{\left(\Lambda_h^{t,i}\right)^{-1}} + \frac{1}{N}\sum_{i=1}^{N}\sum_{h=1}^{H}\mathcal{M}_{i,h}^{\text{on}} \leq \mathcal{O}\left(\sqrt{\frac{d_{\mathbf{c}}\,H^2\log(N/\delta)}{N}}\right). \tag{C.339}$$

It further implies that, with probability at least $1-\delta$,

$$\text{Term (II)} \leq \Gamma_{\mathtt{LMC}}^{\text{on}}\,T\,\max_{t\in[T]}\sum_{h=1}^{H}\mathbb{E}_{\pi^t}\left[\|\phi(s,a)\|_{(\Lambda_h^t)^{-1}}\right] \tag{C.340}$$

$$\overset{\text{(ix)}}{\leq} \mathcal{O}\left(\sqrt{\frac{d_{\mathbf{c}}^3\,H^4\,\log^2(N/\delta)}{N}}\,T\right) \tag{C.341}$$

$$\leq \widetilde{\mathcal{O}}\left(H^2\sqrt{d_{\mathbf{c}}^3\,\log|\mathcal{A}|}\,\sqrt{T}\right), \tag{C.342}$$

where (ix) comes from setting $N = \frac{T}{\log|\mathcal{A}|} = \frac{H^4}{\epsilon^2}$.
Finally, putting everything together, with probability at least $1-\delta$,

$$\mathbb{E}\left[V_1^\star(s_1) - V_1^{\overline{\pi}^T}(s_1)\right] = \frac{1}{T}\left(\text{Term (I)} + \text{Term (II)}\right) = \widetilde{\mathcal{O}}\left(\frac{H^2\sqrt{d_{\mathbf{c}}^3\,\log|\mathcal{A}|}}{\sqrt{T}} + H^2\sqrt{\epsilon}\right). \tag{C.343}$$

This concludes the proof. □

C.5.3 Technical Tools

Lemma 30 (Extended Value Difference)**.** *Given any $\pi, \pi' \in \Delta(\mathcal{A} \mid \mathcal{S}, H)$ and any Q-function $\widehat{Q} \in \mathbb{R}^{H\times|\mathcal{S}|\times|\mathcal{A}|}$, we define $\widehat{V}_h(\cdot) = \mathbb{E}_{a\sim\pi'_h(s,\cdot)}\widehat{Q}_h(\cdot,a)$ for any $h \in [H]$. Then,*

$$\widehat{V}_1(s_1) - V_1^{\pi}(s_1) \tag{C.344}$$

$$= \sum_{h=1}^{H}\mathbb{E}_{s\sim\pi}\left[\left\langle \pi'_h(s,\cdot) - \pi_h(s,\cdot), \widehat{Q}_h(s,\cdot)\right\rangle\right] \tag{C.345}$$

$$+ \sum_{h=1}^{H}\mathbb{E}_{(s,a)\sim\pi}\left[\widehat{Q}_h(s,a) - r_h(s,a) - \sum_{s'\in\mathcal{S}}\mathbb{P}_h(s' \mid s,a)\widehat{V}_{h+1}(s')\right]. \tag{C.346}$$

Lemma 31 (Concentration of Self-Normalized Processes [Abbasi-Yadkori et al., 2011, Theorem 1])**.** *Let $\{x_t\}_{t=1}^{\infty}$ be a real-valued stochastic process with the correspond filtration $\{\mathcal{F}_t\}_{t=0}^{\infty}$ such that x_t is $\mathcal{F}_{t-1}$-measurable, and x_t is conditionally σ-sub-Gaussian for some*

$\sigma > 0$, *i.e.*,

$$\forall \lambda \in \mathbb{R}, \quad \mathbb{E}\left[\exp(\lambda\, x_t) \mid \mathcal{F}_{t-1}\right] = \exp(\lambda^2\, \sigma^2/2)\,. \tag{C.347}$$

Let $\{\phi_t\}_{t=1}^{\infty}$ *be an* $\mathbb{R}^d$*-valued stochastic process such that* ϕ_t *is* $\mathcal{F}_{t-1}$*-measurable. Assume* Λ_0 *is a* $d \times d$ *positive definite matrix, and let* $\Lambda_t = \Lambda_0 + \sum_{i=1}^{t} \phi_i\, \phi_i^\top$*. Then, for any* $\delta > 0$*, with probability at least* $1 - \delta$*, for all* $t \geq 0$*, it holds that*

$$\left\| \sum_{i=1}^{t} \phi_i x_i \right\|^2_{\Lambda_t^{-1}} \leq 2\, \sigma^2 \log\left[\frac{\det(\Lambda_t)^{1/2}\, \det(\Lambda_0)^{-1/2}}{\delta} \right]. \tag{C.348}$$

Lemma 32 (Determinant-Trace Inequality [Abbasi-Yadkori et al., 2011, Lemma 10])**.** *Suppose* $X_1, X_2, \ldots, X_t \in \mathbb{R}^d$ *and for any* $s \in [t]$*,* $\|X\|_2 \leq L$*. Let* $\Lambda_t = \lambda\, I + \sum_{s=1}^{t} X_s\, X_s^\top$ *for some* $\lambda > 0$*. Then, for all* t*, it holds that*

$$\det(\Lambda_t) \leq (\lambda + t\, L^2/d)^d\,. \tag{C.349}$$

Lemma 33 (Abbasi-Yadkori et al. 2011, Lemma 11)**.** *Suppose* $X_1, X_2, \ldots, X_t \in \mathbb{R}^d$ *and for any* $s \in [t]$*,* $\|X\|_2 \leq L$*. Let* $\Lambda_t = \Lambda_0 + \sum_{s=1}^{t} X_s\, X_s^\top$ *and* $\lambda_{\min}(\Lambda_0) \geq \max\{1, L^2\}$*. Then, for all* t*, it hold that*

$$\log\left(\frac{\det(\Lambda_t)}{\det(\Lambda_0)}\right) \leq \sum_{s=1}^{t} \|X_t\|^2_{(\Lambda_t)^{-1}} \leq 2\, \log\left(\frac{\det(\Lambda_t)}{\det(\Lambda_0)}\right). \tag{C.350}$$

C.6 Sample Complexity in the Off-Policy Setting

C.6.1 Covering Number (Proof of Lemma 3)

We first present a bound for the norm of the logit.

Lemma 34. *Consider Algorithm 2 with the* NPG *actor in Algorithm 3. Then, under Assumptions 4 and 5, for all* $(t, h, s, a) \in [T] \times [H] \times \mathcal{S} \times \mathcal{A}$*, it holds that*

$$\left| \left\langle \varphi(s,a), \theta_h^{t,K_t}(s,a) \right\rangle \right| \leq (\bar{\epsilon} + \eta\, H)\, t\,, \tag{C.351}$$

where $\bar{\epsilon}$ *is defined in Lemma 1.*

Proof of Lemma 34. We will prove this by induction. When $t = 0$, since we set $\theta_h^0 = \mathbf{0}$, the statement is trivially true. For $t \geq 1$, assume that the statement stands true for $t - 1$. Since Algorithm 2 optimizes the actor loss up to some errors that are assumed to be bounded, using the triangular inequality, we have that

$$\begin{aligned}
&\left| \left\langle \varphi(s,a), \theta_h^{t,K_t}(s,a) \right\rangle \right| && \text{(C.352)}\\
&\quad = \left| \left\langle \varphi(s,a), \theta_h^{t,K_t} - \widehat{\theta}_h^{t,\star} \right\rangle \right| + \left| \left\langle \varphi(s,a), \widehat{\theta}_h^{t,\star}(s,a) \right\rangle \right| && \text{(C.353)}\\
&\quad \leq \epsilon_{\text{opt}} + \left| \left\langle \varphi(s,a), \widehat{\theta}_h^{t,\star}(s,a) \right\rangle \right| && \text{(C.354)}\\
&\quad \leq \epsilon_{\text{opt}} + \left| \left\langle \varphi(s,a), \widehat{\theta}_h^{t,\star}(s,a) - \theta_h^{t-1,K_{t-1}}(s,a) \right\rangle - \eta\, \widehat{Q}_h^t(s,a) \right| && \text{(C.355)}
\end{aligned}$$

$$+ \left| \left\langle \varphi(s,a), \theta_h^{t-1,K_{t-1}}(s,a) \right\rangle + \eta\, \widehat{Q}_h^t(s,a) \right| \tag{C.356}$$

$$\leq \epsilon_{\text{opt}} + \epsilon_{\text{bias}} + \left| \left\langle \varphi(s,a), \theta_h^{t-1,K_{t-1}}(s,a) \right\rangle + \eta\, \widehat{Q}_h^t(s,a) \right| \tag{C.357}$$

$$\leq \overline{\epsilon} + \left| \left\langle \varphi(s,a), \theta_h^{t-1,K_{t-1}}(s,a) \right\rangle \right| + \left| \eta\, \widehat{Q}_h^t(s,a) \right| \tag{C.358}$$

$$\leq (\overline{\epsilon} + \eta\, H)\, t\,, \tag{C.359}$$

where $\widehat{\theta}_h^{t,\star}$ denotes the optimal actor parameters when optimizing over $\mathcal{D}_{\text{exp}}$ and ρ_{exp}, $\left\langle \varphi(s,a), \theta_h^{t-1,K_{t-1}}(s,a) \right\rangle + \eta\, \widehat{Q}_h^t(s,a)$ is the optimization target in the actor loss of the projected NPG, and the last inequality uses the inductive hypothesis. This concludes the proof. □

Proof of Lemma 3. Consider any $Q, Q' \in \mathcal{Q}$ such that $Q(\cdot,\cdot) = \min\{\langle \phi(\cdot,\cdot), w \rangle\,, H\}^+$ and $Q'(\cdot,\cdot) = \min\{\langle \phi(\cdot,\cdot), w' \rangle\,, H\}^+$. Therefore, we have that

$$\sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} \left| Q(s,a) - Q'(s,a) \right| \leq \sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} \left| \langle \phi(s,a), w - w' \rangle \right| \tag{C.360}$$

$$\leq \sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} \|\phi(s,a)\|\, \|w - w'\| \tag{C.361}$$

$$\leq 2\,\overline{W}\,, \tag{C.362}$$

where the first inequality uses the Cauchy-Schwarz inequality, and the second inequality uses Definition 1, the triangular inequality, and the definition of $\overline{W}$.

Consider any $\pi, \pi' \in \Pi_{\text{lin}}$ such that $\pi(\cdot \mid s) \propto \exp(\langle \phi(s,\cdot), \theta \rangle)$ and $\pi'(\cdot \mid s) \propto \exp(\langle \phi(s,\cdot), \theta' \rangle)$. By invoking Lemma 39 and using Lemma 34, we can observe that for a fixed $s \in \mathcal{S}$,

$$\sup_{a\in\mathcal{A}} \left| \pi(s,a) - \pi'(s,a) \right| \leq \left\| \pi(s,\cdot) - \pi'(s,\cdot) \right\|_1 \leq 2\sqrt{\sup_a \left| \langle \varphi(s,a), \theta - \theta' \rangle \right|} \leq 2\sqrt{2\,\overline{Z}}\,. \tag{C.363}$$

Taking the sup over $\mathcal{S}$, we get that

$$\sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} \left| \pi(s,a) - \pi'(s,a) \right| \leq 2\sqrt{2\,\overline{Z}}\,. \tag{C.364}$$

Therefore, we can bound the logarithm of the covering number of the value function class as follows.

$$\log \mathcal{N}_\Delta(\mathcal{V}) \leq \log \mathcal{N}_{\Delta/2}(\mathcal{Q}) + \log \mathcal{N}_{\Delta/(2H)}(\Pi_{\text{lin}}) \tag{C.365}$$

$$\leq d_{\boldsymbol{c}} \log\left(1 + \frac{4\,\overline{W}}{\Delta}\right) + d_{\boldsymbol{a}} \log\left(1 + \frac{8\,H\sqrt{2\,\overline{Z}}}{\Delta}\right), \tag{C.366}$$

where the first inequality follows from Lemma 36, and the second inequality uses Lemma 38. This concludes the proof. □

C.6.2 Proof of Good Event

Lemma 35. *Consider Algorithm 2 in the off-policy setting with $\lambda = 1$. Then, for any $\delta \in (0,1)$, with probability at least $1-\delta$, it holds that*

$$\left\| \sum_{(s,a)\in\mathcal{D}_h^t} \phi(s,a) \left[\widehat{V}_{h+1}^t(s) - \mathbb{P}_h \widehat{V}_{h+1}^t(s,a) \right] \right\|_{(\Lambda_h^t)^{-1}} \leq C_\delta^{\text{off}} \, H \sqrt{d_{\boldsymbol{c}}} \,, \tag{C.367}$$

where

$$C_\delta^{\text{off}} = 3 \sqrt{\frac{1}{2}\log(T+1) + \log\left(\frac{2\sqrt{2}T}{H}\right) + \log\frac{2}{\delta} + \mathsf{V}} \,, \tag{C.368}$$

$$\mathsf{V} = d_{\boldsymbol{c}} \log\left(1 + \frac{4\,\overline{W} + 4\,H\sqrt{2\,\overline{Z}}}{\Delta}\right) + d_{\boldsymbol{a}} \log\left(1 + \frac{4\,H\sqrt{2\,\overline{Z}}}{\Delta}\right) , \tag{C.369}$$

$$\overline{W} = \frac{16}{3} H \sqrt{d_{\boldsymbol{c}} T} + \sqrt{\frac{2\,d_{\boldsymbol{c}}^3 T}{3\,\zeta\,\delta}} \,, \quad \overline{Z} = (\overline{\epsilon} + \eta H) T \,. \tag{C.370}$$

Proof of Lemma 35. Since $\widehat{V}(\cdot) \in [0.H]$, we can invoke Lemma 37. Then, we have that for any $\Delta > 0$, with probability at least $1-\delta$,

$$\left\| \sum_{(s,a,s')\in\mathcal{D}_h^t} \phi(s,a) \left[\widehat{V}_{h+1}^t(s') - \mathbb{P}_h \widehat{V}_{h+1}^t(s,a) \right] \right\|_{(\Lambda_h^t)^{-1}} \tag{C.371}$$

$$\leq \left(4H^2 \left[\frac{d_{\boldsymbol{c}}}{2} \log\left(\frac{T+\lambda}{\lambda}\right) + d_{\boldsymbol{c}} \log\left(\frac{\mathcal{N}_\Delta(\mathcal{V})}{\Delta}\right) + \log\frac{2}{\delta} \right] + \frac{8\,T^2\Delta^2}{\lambda} \right)^{1/2} \tag{C.372}$$

$$\leq 2H \left[\frac{d_{\boldsymbol{c}}}{2} \log\left(\frac{T+\lambda}{\lambda}\right) + d_{\boldsymbol{c}} \log\left(\frac{\mathcal{N}_\Delta(\mathcal{V})}{\Delta}\right) + \log\frac{2}{\delta} \right]^{1/2} + \frac{2\sqrt{2}T\Delta}{\sqrt{\lambda}} \,. \tag{C.373}$$

Setting $\lambda = 1$, $\Delta = \frac{H}{2\sqrt{2}\,T}$, we have that with probability at least $1-\delta$,

$$\left\| \sum_{(s,a,s')\in\mathcal{D}_h^t} \phi(s,a) \left[\widehat{V}_{h+1}^t(s') - \mathbb{P}_h \widehat{V}_{h+1}^t(s,a) \right] \right\|_{(\Lambda_h^t)^{-1}} \tag{C.374}$$

$$\leq 2\,H\sqrt{d_{\boldsymbol{c}}} \left[\frac{1}{2}\log(T+1) + \log\left(\frac{\mathcal{N}_\Delta(\mathcal{V})}{\frac{H}{2\sqrt{2}T}}\right) + \log\frac{2}{\delta} \right]^{1/2} + H \tag{C.375}$$

$$\leq 3\,H\sqrt{d_{\boldsymbol{c}}} \left[\frac{1}{2}\log(T+1) + \log\left(\frac{2\sqrt{2}T}{H}\right) + \log\frac{2}{\delta} + \mathsf{V} \right]^{1/2} , \tag{C.376}$$

where the last inequality uses Lemma 3 to bound the logarithm of the covering number. This concludes the proof. □

C.6.3 Proof of Theorem 8

We first instantiate Lemma 14 in the off-policy setting. Given the above good event, we have that

$$\Gamma^{\text{off}}_{\texttt{LMC}} \leq \mathcal{O}\left(C^{\text{off}}_{\delta}\, H\, d_{\boldsymbol{c}} \sqrt{\log(1/\delta)}\right) = \widetilde{\mathcal{O}}\left(H \sqrt{d^3_{\boldsymbol{c}}\, \max\{d_{\boldsymbol{c}}, d_{\boldsymbol{a}}\}}\right). \tag{C.377}$$

Proof of Theorem 8. Following the proof of Theorem 7 (Section C.5.2), we can use the same regret decomposition as follows.

$$\sum_{t=1}^{T}\left(V_1^{\star}(s_1) - V_1^{\pi^t}(s_1)\right) \tag{C.378}$$

$$= \underbrace{\sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}_{\pi^\star}\left[\left\langle \pi^{\star}_h(\cdot \mid s) - \pi^t_h(\cdot \mid s), \widehat{Q}^t_h(s,\cdot)\right\rangle\right]}_{\text{(I) policy optimization (actor) error}} + \underbrace{\sum_{t=1}^{T}\sum_{h=1}^{H}\left(\mathbb{E}_{\pi^\star}[\iota^t_h(s,a)] - \mathbb{E}_{\pi^t}[\iota^t_h(s,a)]\right)}_{\text{(II) policy evaluation (critic) error}}. \tag{C.379}$$

Term (I) can be bounded the same way as the proof in Section C.5.2. Hence, it suffices to only bound Term (II).

$$\text{Term (II)} = \sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}_{\pi^\star}[\iota^t_h(s,a)] - \mathbb{E}_{\pi^t}[\iota^t_h(s,a)] \tag{C.380}$$

$$\overset{\text{(i)}}{\leq} -\sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}_{\pi^t}[\iota^t_h(s,a)] \tag{C.381}$$

$$\overset{\text{(ii)}}{\leq} \Gamma^{\text{off}}_{\texttt{LMC}} \sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}_{\pi^t}\left[\left\|\phi(s^t_h, a^t_h)\right\|_{(\Lambda^t_h)^{-1}}\right]. \tag{C.382}$$

(i) and (ii) both follow from Lemma 14, where (i) is based on the optimism guarantee (RHS of Equation (C.89)), while (ii) is based on the error bound (LHS of Equation (C.89)).

Bounding the sum of bonuses. Since $\Gamma^{\text{off}}_{\texttt{LMC}}$ is bounded, it suffices to bound $\mathbb{E}_{\pi^t}\left[\|\phi(s,a)\|_{(\Lambda^t_h)^{-1}}\right]$. We then index each data point in $\mathcal{D}^t_h$ as $\left\{(s^i_h, a^t_h, s^t_{h+1})\right\}_{t\in[T]}$ and get that $\Lambda^t_h = \sum_{t=1}^{T}\phi(s^t_h,a^t_h)\phi(s^t_h,a^t_h)^\top + \lambda I$. Then, we have that

$$\sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}_{\pi^t}\left[\|\phi(s,a)\|_{(\Lambda^t_h)^{-1}}\right] \tag{C.383}$$

$$= \sum_{i=1}^{T}\sum_{h=1}^{H}\left\|\phi(s^t_h,a^t_h)\right\|_{\left(\Lambda^{t,i}_h\right)^{-1}} + \sum_{t=1}^{T}\sum_{h=1}^{H}\underbrace{\mathbb{E}_{\substack{s_h\sim\mathbb{P}(\cdot|s^t_{h-1}a^t_{h-1})\\ a_h\sim\pi^t_h(\cdot|s_h)}}\left[\|\phi(s_h,a_h)\|_{\left(\Lambda^t_h\right)^{-1}}\right] - \left\|\phi(s^t_h,a^t_h)\right\|_{\left(\Lambda^t_h\right)^{-1}}}_{:=\mathcal{M}^{\text{off}}_{t,h}}. \tag{C.384}$$

Applying the elliptical potential lemma. For the first term of Equation (C.384), we have that

$$\sum_{t=1}^{T}\sum_{h=1}^{H}\left\|\phi(s_h^t,a_h^t)\right\|_{\left(\Lambda_h^t\right)^{-1}} = \sum_{h=1}^{H}\sum_{t=1}^{T}\left\|\phi(s_h^t,a_h^t)\right\|_{\left(\Lambda_h^t\right)^{-1}} \tag{C.385}$$

$$\overset{\text{(iii)}}{\leq} \sum_{h=1}^{H}\sqrt{T}\left(\sum_{t=1}^{T}\left\|\phi(s_h^t,a_h^t)\right\|^2_{\left(\Lambda_h^t\right)^{-1}}\right)^{1/2} \tag{C.386}$$

$$\overset{\text{(iv)}}{\leq} \mathcal{O}\left(\sqrt{d_{\boldsymbol{c}}\,H^2\,T\log(T/\delta)}\right). \tag{C.387}$$

(iii) applies the Cauchy-Schwarz inequality, and (iv) follows the elliptical potential argument from Lemma 33.

A martingale difference sequence. For the second term of Equation (C.384), since $\left\{\mathcal{M}_{t,h}^{\text{off}}\right\}_{(t,h)\in[T]\times[H]}$ forms a martingale sequence adapted to the filtration,

$$\mathcal{F}_{t,h}^{\text{off}} = \left\{(s_\tau^i,a_\tau^i)\right\}_{(i,\tau)\in[t-1]\times[H]} \cup \left\{(s_\tau^t,a_\tau^t)\right\}_{\tau\in[h-1]}, \tag{C.388}$$

such that $\mathbb{E}\left[\mathcal{M}_{t,h}^{\text{off}} \mid \mathcal{F}_{t,h}^{\text{off}}\right] = 0$. Since $|\mathcal{M}_{i,h}^{\text{off}}| \leq 1$, we can apply the Azuma–Hoeffding inequality and obtain that

$$\Pr\left(\sum_{t=1}^{T}\sum_{h=1}^{H}\mathcal{M}_{t,h}^{\text{off}} \geq m\right) \geq \exp\left(\frac{-m^2}{2\,H\,T}\right). \tag{C.389}$$

Setting $m = \sqrt{2\,H\,T\,\log(1/\delta)}$, with probability at least $1-\delta$, it holds that

$$\sum_{t=1}^{T}\sum_{h=1}^{H}\mathcal{M}_{t,h}^{\text{off}} \leq \sqrt{2\,H\,T\,\log(1/\delta)} \leq \mathcal{O}\left(\sqrt{H\,T\,\log(1/\delta)}\right). \tag{C.390}$$

Therefore, we have that

$$\sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}_{\pi^t}\left[\|\phi(s,a)\|_{(\Lambda_h^t)^{-1}}\right] = \sum_{t=1}^{T}\sum_{h=1}^{H}\left\|\phi(s_h^i,a_h^i)\right\|_{\left(\Lambda_h^{t,i}\right)^{-1}} + \sum_{t=1}^{T}\sum_{h=1}^{H}\mathcal{M}_{i,h}^{\text{off}} \leq \mathcal{O}\left(\sqrt{d_{\boldsymbol{c}}\,H^2\,T\log(T/\delta)}\right). \tag{C.391}$$

It further implies that, with probability at least $1-\delta$,

$$\text{Term (II)} \leq \Gamma_{\texttt{LMC}}^{\text{off}}\sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}_{\pi^t}\left[\|\phi(s,a)\|_{(\Lambda_h^t)^{-1}}\right] \leq \widetilde{\mathcal{O}}\left(\sqrt{d_{\boldsymbol{c}}^3\,\max\{d_{\boldsymbol{c}},d_{\boldsymbol{a}}\}\,H^4\,T}\right). \tag{C.392}$$

Putting everything together. Therefore, we have that with probability at least $1-\delta$,

$$\mathbb{E}\left[V_1^\star(s_1) - V_1^{\overline{\pi}^T}(s_1)\right] = \frac{1}{T}\left(\text{Term (I)} + \text{Term (II)}\right) = \widetilde{\mathcal{O}}\left(\frac{H^2\sqrt{d_{\boldsymbol{c}}^3\,\max\{d_{\boldsymbol{c}},d_{\boldsymbol{a}}\}\,\log|\mathcal{A}|}}{\sqrt{T}} + H^2\sqrt{\overline{\epsilon}}\right). \tag{C.393}$$

This concludes the proof. □

C.6.4 Technical Tools

Lemma 36 (Zhong and Zhang, 2023, Lemma B.1)**.** *Consider the value function class* $\mathcal{V} = \{\langle Q(\cdot,\cdot), \widehat{\pi}(\cdot \mid \cdot)\rangle_{\mathcal{A}} \mid Q \in \mathcal{Q}, \widehat{\pi} \in \Pi\}$*. Then, it holds that*

$$\mathcal{N}_{\Delta}(\mathcal{V}) \leq \mathcal{N}_{\Delta/2}(\mathcal{Q}) \cdot \mathcal{N}_{\Delta/(2\,H)}(\Pi)\,. \tag{C.394}$$

Lemma 37 (Value-Aware Uniform Concentration [Jin et al., 2020, Lemma D.4])**.** *Let* $\{s_t\}_{t=1}^{\infty}$ *be a stochastic process on the state space* $\mathcal{S}$ *with the correspond filtration* $\{\mathcal{F}_t\}_{t=0}^{\infty}$ *such that* s_t *is* $\mathcal{F}_{t-1}$*-measurable. Let* $\{\phi_t\}_{t=1}^{\infty}$ *be an* $\mathbb{R}^d$*-valued stochastic process such that* ϕ_t *is* $\mathcal{F}_{t-1}$*-measurable, and* $\|\phi_t\| \leq 1$*. Let* $\Lambda_t = I + \sum_{s=1}^{t} \phi_s\,\phi_s^{\top}$*. Assume* $\mathcal{V}$ *is a value function class such that* $\sup_{s\in\mathcal{S}} |V(s)| \leq H$*. Then, for any* $\delta > 0$*, with probability at least* $1-\delta$*, for all* $t \geq 0$ *and any* $V \in \mathcal{V}$*, it holds that*

$$\left\| \sum_{i=1}^{t} \phi_i \left\{ V(s_i) - \mathbb{E}\left[V(s_i) \mid \mathcal{F}_{i-1}\right] \right\} \right\|_{\Lambda_t^{-1}}^2 \leq 4\,H^2 \left[\frac{d}{2} \log\left(\frac{t+\lambda}{\lambda} \right) + \log\left(\frac{\mathcal{N}_{\Delta}}{\delta} \right) \right] + \frac{8\,t^2\,\Delta^2}{\lambda}\,, \tag{C.395}$$

where $\mathcal{N}_{\Delta}$ *represents the* Δ*-covering number of* $\mathcal{V}$ *with the distance measured by* $dist(V, V') = \sup_{s\in\mathcal{S}} |V(s) - V'(s)|$*.*

Lemma 38 (Covering Number of Euclidean Ball)**.** *For any* $\Delta > 0$*, the* Δ*-covering number,* $\mathcal{N}_{\Delta}$*, of the Euclidean ball of radius* $B > 0$ *in* $\mathbb{R}^d$ *satisfies that*

$$\mathcal{N}_{\Delta} \leq \left(1 + \frac{2B}{\Delta} \right)^d . \tag{C.396}$$

Lemma 39 (Zhong and Zhang 2023, Lemma B.3)**.** *For* $\pi, \pi' \in \Delta(\mathcal{A})$ *and* $Z, Z' : \mathcal{A} \to \mathbb{R}^+$*, if* $\pi(\cdot) \propto \exp(Z(\cdot))$ *and* $\pi'(\cdot) \propto \exp(Z'(\cdot))$*, then it holds that*

$$\|\pi - \pi'\|_1 \leq 2\sqrt{\|Z - Z'\|_{\infty}}\,. \tag{C.397}$$

C.7 Experiments

In this section, we evaluate the performance of our proposed algorithm with other methods on various benchmarks. In Section C.7.1, we test our proposed algorithm in two specific environments of linear MDPs. In Section C.7.2, we further conduct some ablation studies in the same two environments. The Atari results are presented in Section 4.7.2; their implementation details, environment setup, and hyperparameters are provided in Section C.7.3.

C.7.1 Experiments in Linear MDPs

First, we test our proposed algorithm in the randomly generated linear MDPs (Random MDP) and the linear MDP version of the Deep Sea [Osband et al., 2019].

Environment Setup

Our experimental setup is an extension of Ishfaq et al. [2024a]. In particular, we extend the prior off-policy setting to test our proposed algorithm in the linear MDP version of the Deep Sea [Osband et al., 2019] and the Random MDP. In both experiments, we use the linear MDP features as the policy features (i.e., $\phi = \varphi$), and we set $d \coloneqq d_{\boldsymbol{c}} = d_{\boldsymbol{a}}$ to represent the feature dimension for both the actor and the critic parameters.

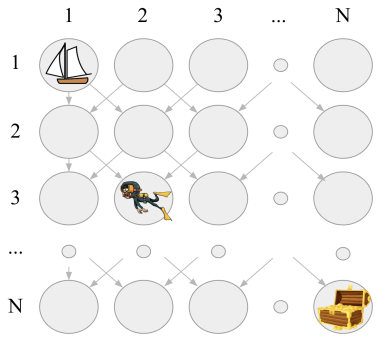


Figure C.1: Example of the Deep Sea environment from Osband et al. [2019].

For the Random MDP environment, we consider 15 states and 5 actions. For each state $s \in \mathcal{S}$, we generate $\psi_h(s) \in \mathbb{R}^d$ uniformly at random in $[0, 1]$ and construct tile coded features. The agent always starts from state 0, receiving a small reward of 0.1 upon taking action 0, and obtains the maximum reward when reaching the final state and taking action 1. All other state-action pairs yield zero reward. Given this reward function and a randomly generated transition kernel, we solve for ψ_h and υ_h via minimizing least square errors following Definition 1 and use the corresponding $\mathbb{P}_h$ and r_h to set up the environment.

For the Deep Sea environment, we use a $N \times N$ grid with $N = 10$ where the agent always starts at $(1, 1)$ and can move either bottom-right or bottom-left, receiving rewards of 0 and $-0.01/N$ respectively. Reaching the bottom-right corner yields a reward of 1. Furthermore, we generate the actor and critic features by projecting each state-action pair uniformly between $[0, d-1]$, which recovers one-hot encoded features when $d = |\mathcal{S}| \times |\mathcal{A}|$. Given the true transition probabilities and rewards, we solve for ψ_h and υ_h via minimizing least square errors following Definition 1 and use the corresponding $\mathbb{P}_h$ and r_h to ensure the linearity of the MDP.

Coreset Construction

To implement our proposed algorithm, we need to conduct the experimental design to obtain $\mathcal{D}_{\text{exp}}$ and ρ_{exp}. For this, we follow the offline G-Experimental design outlined in Algorithm 6 to construct a coreset. In particular, in each iteration, this greedy iterative algorithm traverses the entire state-action space and adds a data point to the coreset that has the highest marginal gain $g(s, a) = \|\varphi(s, a)\|_{G^{-1}}$. For a specific threshold ϵ_G, the algorithm only terminates when $g_{\max} = \max_{s,a \in (\mathcal{S} \times \mathcal{A})} g(s, a) \leq \epsilon_G$, hence giving us direct control over $\sup_{(s,a) \in \mathcal{S} \times \mathcal{A}} \|\varphi(s, a)\|_{G^{-1}}$. In practice, we find that it often selects too many data points, so we cap the coreset at 80% of the total data.

Algorithm 6: Coreset Construction Using G-Experimental Design

Input: features $\varphi : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}^{d_{\boldsymbol{a}}}$, threshold $\epsilon_G \in \mathbb{R}$
Initialize: $G = I_{d_{\boldsymbol{a}} \times d_{\boldsymbol{a}}}$, $\mathcal{D}_{\text{exp}} = \emptyset$, $g_{\max} = \infty$
while $g_{\max} > \epsilon_G$ **do**
  $g_{\max} = 0$
  for $(s, a) \in \mathcal{S} \times \mathcal{A}$ **do**
    $g(s, a) = \|\varphi(s, a)\|_{G^{-1}}$
    if $g_{\max} < g(s, a)$ **then**
      $(s^\star, a^\star) = (s, a)$
      $g_{\max} = g(s, a)$
    end if
    $\mathcal{D}_{\text{exp}} = \mathcal{D}_{\text{exp}} \cup \{(s^\star, a^\star)\}$
    $G = G + \varphi(s^\star, a^\star)\, \varphi(s^\star, a^\star)^\top$
  end for
end while

Algorithms and Hyperparameters

We denote by LMC-NPG-EXP our proposed algorithm with an explicit log-linear policy parameterization that uses LMC for policy evaluation and projected NPG for policy optimization over the obtained coreset. We denote by LMC-NPG-IMP an idealized variant of NPG that does not have an explicit policy parameterization and maintains an implicit policy by storing all parameterized Q functions (and hence requires significantly more memory). As a baseline, we also consider the value-based algorithm LMC [Ishfaq et al., 2024a].

In Table C.2, we list the hyperparameters used across all experiments. For log-linear policies, the actor loss in Equation (4.13) admits a closed-form solution, allowing us to avoid tuning of the actor learning rate $\alpha_{\boldsymbol{a}}$ and the number of actor updates K_t, by minimizing the objective exactly. In general, for non-linear models, inexact optimization (e.g., stochastic gradient descent) is usually required to optimize the actor loss.

Hyperparameter	**LMC**	**LMC-NPG-IMP**	**LMC-NPG-EXP**
Policy Optimization Learning Rate (η)	✗	$[0.1, 1, 10, 100]$	$[0.1, 1, 10, 100]$
Inverse Temperature (ζ^{-1})	$[10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}]$		
Number of Critic Updates (J_t)	100		
Critic Learning Rate ($\alpha_{\boldsymbol{c}}$)	$[10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}]$		
Number of Episodes (T)	600		
Horizon Length (H)	100		

Table C.2: Hyperparameter search space for our experiments in linear MDPs.

Experimental Results

Following the protocol of Ishfaq et al. [2024a], each algorithm is run with 20 random seeds. We sweep the hyperparameters as shown in Table C.2 and report the best performance with 95% confidence intervals.

For the Random MDP, Figure 4.1 indicates that LMC-NPG-EXP closely matches LMC-NPG-IMP while outperforming the value-based baseline, LMC. For the linear MDP version of Deep

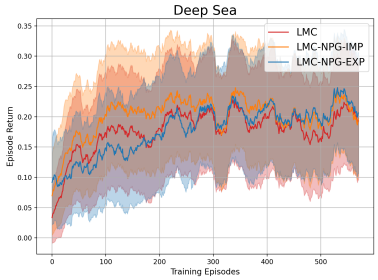


Figure C.2: Comparison of `LMC-NPG-EXP` (our proposed framework), `LMC-NPG-IMP` (memory-intensive variant), and `LMC` (value-based baseline) in the Deep Sea environment.

Sea, Figure C.2 showcases that `LMC-NPG-EXP` can achieve comparable performance with `LMC-NPG-IMP` and `LMC`.

C.7.2 Ablation Studies

Ablation on Exploration

To study the impact of the exploration mechanism, `LMC`, in our proposed algorithm, we perform an ablation in the linear MDP variant of Deep Sea. For the baseline without exploration, we consider the same algorithm design as `LMC-NPG-EXP` and simply do not inject any noise into the LMC update. The results in Figure C.3 indicate that when the feature dimensions of the critic and the actor are relatively small, the exploration mechanism is crucial for our proposed algorithm to achieve decent performance.

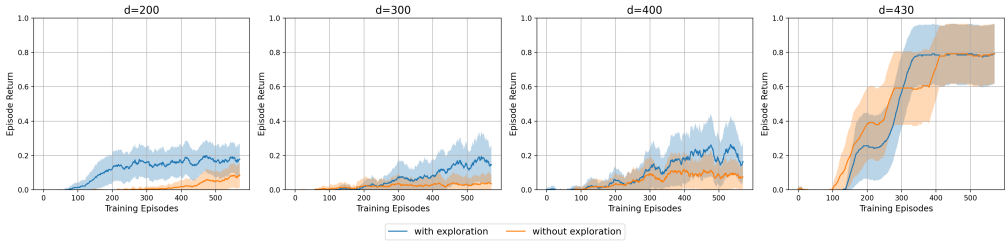


Figure C.3: Ablation of the exploration mechanism for `LMc-NPG-EXP`.

Ablation on Feature Dimensions

We also study the effect of the feature dimensions. We use the same feature for the MDP environment and the policy (i.e., $\phi = \varphi$), and we denote that $d := d_{\boldsymbol{c}} = d_{\boldsymbol{a}}$. The results in Figure C.4 show that larger feature dimensions d for both the actor and the critic lead to greater performance of the proposed algorithm.

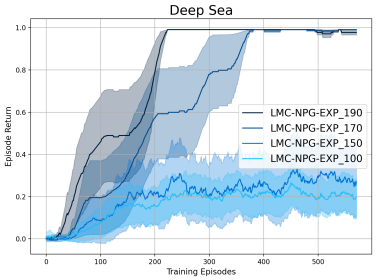


Figure C.4: Effect of feature dimension d in Deep Sea.

Sensitivity to Inverse Temperature (ζ^{-1})

We study the sensitivity of our algorithm to the inverse temperature hyperparameter, ζ^{-1}, in the Deep Sea environment. As shown Figure C.5, we observe relatively robust performance for different choices of this hyperparameter across a range of different feature dimensions.

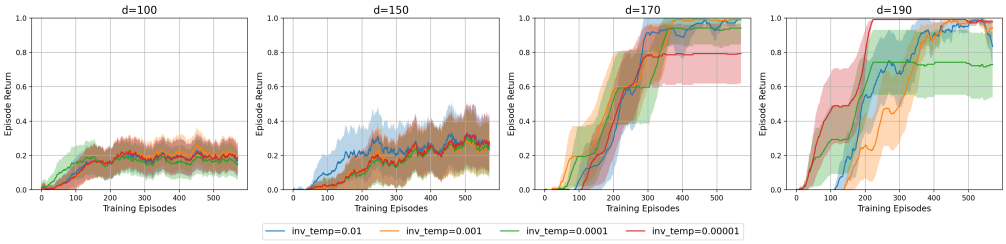


Figure C.5: `LMC-NPG-EXP` exhibits robustness to different choices of ζ^{-1}.

Sensitivity to the Number of Critic Samples (M)

We further study the effect of the number of critic samples, M, in the Deep Sea environment. We vary M across a range of different feature dimensions. Similarly, as shown in Figure C.6, we find that the performance of our proposed algorithm remains relatively stable across the tested range of different feature dimensions.

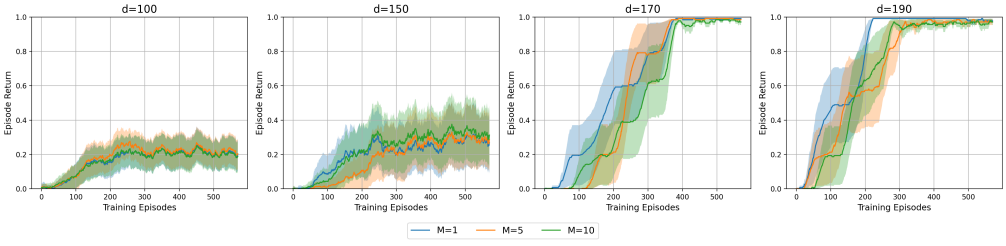


Figure C.6: `LMC-NPG-EXP` exhibits robustness to different choices of M.

C.7.3 Atari Experimental Details

This subsection provides the implementation and experimental details for the Atari results in Section 4.7.2.

Extension to Deep RL Applications

Compared to our theoretical results in the finite-horizon linear MDP setting, the Atari benchmark requires handling discounted problems with complex nonlinear function approximation. Consequently, starting from our theoretically principled algorithm, we make three mild changes and follow the same protocol as the practical version of `LMC` in Ishfaq et al. [2024a]. Since we are not in the finite-horizon setting, we compute the value functions and Q functions in a forward fashion ($h = 1, \ldots$) rather than backwards ($h = H, \ldots, 1$), which changes how the critic loss is formed in Algorithm 4. Moreover, instead of using stochastic gradient descent to optimize the `LMC` critic loss, we follow Ishfaq et al. [2024a] and use the practical version of `LMC` that integrates Adam. Finally, rather than relying on the experimental design, at iteration t, we roll out the current policy to store state–action pairs in a buffer $\mathcal{D}^t$ and minimize the actor objective over (a subset of) it. Following such changes, we can extend our proposed algorithm beyond the linear function approximation and empirically test its performance against other commonly used algorithms.

Environment Setups and Hyperparameters

In the Atari experiments, we use the following setups. In the on-policy setting, we adopt the recommended hyperparameters from Raffin [2020] for `PPO` and our algorithm. In the off-policy setting, following prior work [Tomar et al., 2020], we use the default hyperparameters from stable baselines [Raffin et al., 2021] for `DQN` and our algorithm. These setups are motivated by two considerations. First, we aim to evaluate the effect of different objectives without performing an extensive hyperparameter search. Second, the CNN-based actor and the critic architectures make large grid searches over multiple hyperparameters (e.g., framestack, λ in GAE, horizon length, and discount factor) computationally prohibitive. A complete list of hyperparameters used in the on-policy and off-policy Atari experiments is provided in Tables C.3 and C.4. Additionally, for the policy optimization learning rate η in our algorithm, we perform a grid search over $\{0.01, 0.1, 1.0\}$.

Hyperparameter	**LMC-NPG-EXP**	**PPO**
Reward normalization	✗	✗
Observation normalization	✗	✗
Orthogonal weight initialization	✓	✓
Value function clipping	✗	✗
Gradient clipping	✗	✓
Probability ratio clipping	✗	✓
Clip range	✗	0.1
Entropy coefficient	0	0.01
Number of inner loop updates (m)	5	4
Adam step-size	3×10^{-4}	2.5×10^{-4}
Value Function Coefficient	✗	0.5
Minibatch size	256	
Framestack	4	
Number of environment copies	8	
GAE (λ)	0.95	
Horizon (T)	128	
Discount factor	0.99	
Total number of timesteps	10^7	
Number of runs for plot averages	5	
Confidence interval for plot runs	$\sim 95\%$	

Table C.3: Hyperparameters for the Atari experiments in the on-policy setting.

Hyperparameter	**LMC-NPG-EXP**	**DQN**
Reward normalization	✗	✗
Observation normalization	✗	✗
Orthogonal weight initialization	✗	✗
Value function clipping	✗	✗
Gradient clipping	✗	✗
Probability ratio clipping	✗	✗
Exploration	LMC	ϵ-greedy
Adam step-size	3×10^{-4}	
Buffer size	10^6	
Minibatch size	256	
Framestack	4	
Number of environment copies	8	
Discount factor	0.99	
Total number of timesteps	10^7	
Number of runs for plot averages	5	
Confidence interval for plot runs	$\sim 95\%$	

Table C.4: Hyperparameters for the Atari experiments in the off-policy setting.