Algorithmic Foundations of Policy Optimization in Reinforcement Learning, Multi-agent Systems, and AI Alignment

Submitted in partial fulfillment of the requirements for

the degree of

Doctor of Philosophy

in

Electrical and Computer Engineering

Shicong Cen

B.S., Information and Computing Science, Peking University

Carnegie Mellon University

Pittsburgh, PA

August 2024

©Shicong Cen, 2024

All Rights Reserved

Acknowledgments

The research in this thesis is supported in part by the grants ONR N00014-18-1-2142 and N00014-19-1-2404, ARO W911NF-18-1-0303, NSF CCF-1806154, CCF-1901199, CCF-2007911, CCF-2106778, DMS-2134080 and CNS-214821, as well as J.P. Morgan AI Research Fellowship, Wei Shen and Xuehong Zhang Presidential Fellowship, Boeing Fellowship, and Nicholas Minnici Dean's Graduate Fellowship in Electrical and Computer Engineering at Carnegie Mellon University.

First and foremost, I would like express my deepest gratitude to my advisor and thesis committee chair, Prof. Yuejie Chi, for her unwavering support, guidance, and encouragement throughout the course of my research. Yuejie provided me with the intellectual freedom to explore my ideas while offering the structure and wisdom needed to keep me on track. I constantly get inspired by Yuejie's passion for research, and feel truly fortunate to have had the opportunity to work under her mentorship. Beyond academic guidance, Yuejie has provided invaluable advice on my personal career planning and helped honing my presentation skills, which have prepared me for future challenges.

I would also like to extend my heartfelt thanks to the other members of my thesis committee, Prof. Guannan Qu, Prof. Maryam Fazel, and Prof. Dale Schuurmans, for their valuable insights, constructive feedback, and continuous support throughout this process. Their expertise and thoughtful suggestions have significantly enhanced the quality of this thesis, and I deeply appreciate the time and effort they have invested in reviewing my work.

I am deeply grateful to Dr. Lin Xiao and Prof. Simon S. Du from Meta's FAIR team, as well as Dr. Jincheng Mei and Prof. Bo Dai from Google DeepMind, for their exceptional guidance during my internships. Their thoughtful advice and encouragement greatly contributed to the progress and success of my projects, and I am thankful for the opportunity to learn from them. I also extend my sincere thanks to Prof. Yuxin Chen and Prof. Yuting Wei for their invaluable guidance and supervision throughout much of this thesis, particularly for their assistance in shaping the first paper draft during my Ph.D. journey.

Moreover, I would like to acknowledge my research groupmates at CMU — Harlin Lee, Vince Monardo, Tian Tong, Boyue Li, Laixi Shi, Maxime Ferreira Da Costa, Zhize Li, Pedro Valdeira, Jiin Woo, Harry Dong, Lingjing Kong, Zixin Wen, He Wang, Xingyu Xu and Tong Yang — for their inspiration and support throughout this journey. I am equally grateful to Prof. Yuxin Chen and Prof. Yuting Wei's group members for jointly hosting reading groups with insightful discussions. My sincere thanks also go to my wonderful collaborators, Chen Cheng, Wenhao Zhan, Fan Chen, and Ruicheng Ao, for their key contributions to the technical breakthroughs in this thesis.

Finally, I extend my heartfelt thanks to my parents for their unconditional love and support over the years.

Shicong Cen

Abstract

Reinforcement learning (RL) aims to solve various tasks by modeling them as learning and sequential decision-making problems within an unknown environment. The empirical success of contemporary RL applications largely owes to policy optimization methods, which seek to optimize a parameterized policy to maximize the value function induced by executing the policy. Despite their widespread practical adoption, the theoretical foundations of these approaches remain limited, particularly due to the intrinsic non-concavity of the objective and non-stationarity issues in multi-agent settings. Although substantial progress has been made in understanding the computational feasibility of policy optimization methods, existing results still fall short of explicitly characterizing the iteration complexity across a broad range of RL scenarios.

The first part of this thesis contributes to the algorithmic foundations of policy optimization by investigating non-asymptotic convergence guarantees and improving dependencies on key problem parameters. For tabular single-agent RL, this thesis examines the natural policy gradient (NPG) method with entropy regularization, demonstrating that the method provably converges to the optimal regularized policy at a dimension-free linear rate. Beyond entropy regularization, this thesis develops a novel policy optimization method with the same linear convergence rate that accommodates various choices of regularizers, even those lacking strong convexity and smoothness.

The second part of this thesis extends the study to various multi-agent systems, aiming to provide better iteration complexity bounds for finding approximate Nash equilibrium (NE). For two-player zero-sum matrix games, this thesis introduces novel extra-gradient policy optimization methods that provably converge to the regularized NE at a dimension-free linear rate, which are further generalized to two-player zero-sum Markov games and multi-player zero-sum polymatrix games. Notably, the analysis offers last-iterate convergence guarantees without the need of introducing additional uniqueness assumption and unknown constants, which are typical in existing results. For multi-player potential games, this thesis establishes new iteration complexity bounds for independent entropy-regularized NPG in finding a regularized NE, scaling sub-linearly with the number of agents and independently of the action space size.

The final part of this thesis focuses on the statistical aspects of policy optimization, proposing a unified algorithmic framework that imbues policy optimization methods with principled optimism or pessimism under uncertainty. Specifically, this thesis explores reinforcement learning from human feedback (RLHF) to align large language models (LLMs) with human preferences, developing novel, practically implementable policy optimization methods that regularize the maximum-likelihood estimate of the reward function with the corresponding value function. This approach circumvents the intractable construction of confidence intervals typical in standard implementations of optimism/pessimism principles, and shares a simpler RLHF pipeline akin to direct preference optimization by directly optimizing the policy.

Contents

1 Introduction **1**
- 1.1 Efficient policy optimization for single-agent RL . . . 2
  - 1.1.1 Entropy-regularized RL . . . 2
  - 1.1.2 General regularized RL . . . 5
- 1.2 Efficient policy optimization for multi-agent systems . . . 8
  - 1.2.1 Two-player zero-sum matrix games . . . 8
  - 1.2.2 Two-player zero-sum Markov games . . . 11
  - 1.2.3 Multi-player zero-sum polymatrix games . . . 14
  - 1.2.4 Multi-player potential games . . . 16
- 1.3 Principled policy optimization for AI alignment . . . 17
- 1.4 Related works . . . 19
  - 1.4.1 Single-agent RL . . . 19
  - 1.4.2 Multi-agent systems . . . 20
  - 1.4.3 AI alignment . . . 22
- 1.5 Thesis organization and notation . . . 23

I Policy optimization for single-agent RL **25**

2 Entropy-regularized Natural Policy Gradient Method **26**
- 2.1 Model and algorithms . . . 26
  - 2.1.1 Problem settings . . . 26
  - 2.1.2 Algorithm: NPG methods with entropy regularization . . . 28
- 2.2 Main results . . . 29
  - 2.2.1 Exact entropy-regularized NPG methods . . . 29
  - 2.2.2 Approximate entropy-regularized NPG methods . . . 32
- 2.3 Discussion . . . 33

3 Generalized Policy Mirror Descent Method **35**
- 3.1 Model and algorithms . . . 35
  - 3.1.1 Problem settings . . . 35
  - 3.1.2 Algorithm: generalized policy mirror descent . . . 36
- 3.2 Main results . . . 39
  - 3.2.1 Convergence of exact GPMD . . . 39
  - 3.2.2 Convergence of approximate GPMD . . . 40
- 3.3 Discussion . . . 42

II Policy optimization for multi-agent Systems **45**

4 Two-player Zero-sum Matrix Games **46**
4.1 Background and problem formulation . . . 46
4.2 Proposed extragradient methods: PU and OMWU . . . 47
4.3 Last-iterate linear convergence guarantees . . . 49
4.4 Discussion . . . 52

5 Two-player Zero-sum Markov Games **53**
5.1 Algorithm and theory: the infinite-horizon setting . . . 53
5.1.1 Problem formulation . . . 53
5.1.2 Single-loop algorithm design . . . 55
5.1.3 Theoretical guarantees . . . 56
5.2 Algorithm and theory: the finite-horizon setting . . . 58
5.3 Discussion . . . 59

6 Multi-player Zero-sum Polymatrix Games **60**
6.1 Preliminaries . . . 60
6.2 Performance guarantees of single-timescale OMWU . . . 62
6.2.1 Performance guarantees without delays . . . 62
6.2.2 Performance guarantees under random delays . . . 64
6.3 Performance guarantees of two-timescale OMWU . . . 65
6.3.1 Performance guarantees under constant and known delays . . . 66
6.3.2 Performance guarantees with permuted bounded delays . . . 66
6.4 Discussion . . . 67

7 Multi-player Potential Games **69**
7.1 Potential games with entropy regularization . . . 69
7.1.1 Potential games . . . 69
7.1.2 Entropy-regularized potential games . . . 70
7.2 Finite-time global convergence of independent natural policy gradient methods . . . 71
7.2.1 Independent natural policy gradient method . . . 71
7.2.2 Finite-time global convergence . . . 72
7.3 Discussion . . . 74

III Principled AI alignment **75**

8 Reinforcement Learning from Human Feedback **76**
8.1 Preliminaries . . . 76
8.2 Value-incentivized preference optimization . . . 77
8.2.1 General framework . . . 77
8.2.2 Online RLHF: algorithm and theory . . . 79
8.2.3 Offline RLHF: algorithm and theory . . . 81
8.3 Experiments . . . 83
8.3.1 Synthetic multi-armed bandits . . . 83
8.4 Discussion . . . 83

A Proofs for Chapter 2 **85**
A.1 Analysis . . . 85
A.1.1 Main pillars for the convergence analysis . . . 85
A.1.2 Analysis of exact entropy-regularized NPG methods . . . 87
A.1.3 Analysis of approximate entropy-regularized NPG methods . . . 90
A.2 Preliminaries . . . 91
A.2.1 Derivation of entropy-regularized NPG methods . . . 91
A.2.2 Basic facts about the function $\log(\|\exp(\theta)\|_1)$ . . . 92
A.3 Proof for key lemmas . . . 92
A.3.1 Proof of Lemma 3 . . . 92
A.3.2 Proof of Lemma 4 . . . 94
A.3.3 Proof of Lemma 5 . . . 95
A.3.4 Proof of Lemma 6 . . . 96
A.3.5 Proof of Lemma 7 . . . 97
A.4 Proof for approximate entropy-regularized NPG (Theorem 2) . . . 100

B Proofs for Chapter 3 **104**
B.1 Analysis for exact GPMD (Theorem 3) . . . 104
B.1.1 Preparation: basic facts . . . 104
B.1.2 Proof of Theorem 3 . . . 105
B.2 Proof of key lemmas . . . 107
B.2.1 Proof of Lemma 1 . . . 107
B.2.2 Proof of Lemma 8 . . . 108
B.2.3 Proof of Lemma 9 . . . 111
B.2.4 Proof of Lemma 10 . . . 112

C Proofs for Chapter 4 **114**
C.1 Analysis for entropy-regularized matrix games . . . 114
C.1.1 Proof of Proposition 1 . . . 115
C.1.2 Proof of Theorem 5 . . . 115
C.2 Proof of auxiliary lemmas . . . 124
C.2.1 Proof of Lemma 13 . . . 124
C.2.2 Proof of Lemma 14 . . . 124
C.2.3 Proof of Lemma 16 . . . 125

D Proofs for Chapter 5 **127**
D.1 Analysis for the infinite-horizon setting . . . 127
D.2 Analysis for the finite-horizon setting . . . 132
D.3 Proof of key lemmas for the infinite-horizon setting . . . 135
D.3.1 Proof of Lemma 17 . . . 135
D.3.2 Proof of Lemma 18 . . . 138
D.3.3 Proof of Lemma 19 . . . 140
D.3.4 Proof of Lemma 20 . . . 144
D.3.5 Proof of Lemma 21 . . . 145
D.3.6 Proof of Lemma 22 . . . 146
D.4 Proof of key lemmas for the finite-horizon setting . . . 147
D.4.1 Proof of Lemma 25 . . . 147
D.4.2 Proof of Lemma 26 . . . 150

D.5 Proof of auxiliary lemmas . . . 152
D.5.1 Proof of Lemma 27 . . . 152
D.5.2 Proof of Lemma 28 . . . 153
D.5.3 Proof of Lemma 29 . . . 155
D.5.4 Proof of Lemma 30 . . . 156
D.5.5 Proof of Lemma 32 . . . 157
D.5.6 Proof of Lemma 33 . . . 157

E Proofs for Chapter 6 **160**
E.1 Proof for single-timescale OMWU (Section 6.2) . . . 160
E.1.1 Proof of Theorem 8 . . . 160
E.1.2 Proof of Theorem 9 . . . 163
E.2 Proof for two-timescale OMWU (Section 6.3) . . . 170
E.2.1 Proof of Theorem 10 . . . 170
E.2.2 Proof of Theorem 11 . . . 176
E.3 Proof of auxiliary lemmas . . . 183
E.3.1 Proof of Lemma 35 . . . 183
E.3.2 Proof of Lemma 36 . . . 184
E.3.3 Proof of Lemma 37 . . . 185
E.3.4 Proof of Lemma 38 . . . 186
E.3.5 Proof of Lemma 39 . . . 187
E.3.6 Proof of Lemma 40 . . . 188
E.3.7 Proof of Lemma 41 . . . 189
E.3.8 Proof of Lemma 42 . . . 191

F Proofs for Chapter 7 **194**
F.1 Proof of Theorem 12 . . . 194
F.1.1 Step 1: quantify the policy improvement . . . 194
F.1.2 Step 2: introduce the auxiliary sequence . . . 195
F.1.3 Step 3: bound the gap . . . 195
F.2 Proof of Lemma 45 . . . 196
F.3 Proof of Lemma 46 . . . 196
F.3.1 Proof of (F.5a) . . . 197
F.3.2 Proof of (F.5b) . . . 199
F.3.3 Proof of Corollary 1 . . . 200

G Proofs for Chapter 8 **201**
G.1 Analysis for the online setting . . . 201
G.1.1 Proof of Theorem 13 . . . 201
G.1.2 Proof of Lemma 48 . . . 205
G.1.3 Proof of Lemma 50 . . . 206
G.2 Analysis for the offline setting . . . 207
G.2.1 Proof of Lemma 2 . . . 207
G.2.2 Proof of Theorem 14 . . . 207

List of Tables

1.1 The iteration complexities of NPG methods to reach ε-accuracy *in terms of optimization error*, where the unregularized (resp. regularized) version is given by (2.12) (cf. (2.14)) with η the learning rate. We assume exact gradient evaluation and softmax parameterization, and hide the dependencies that are logarithmic on problem parameters. Here, ε-accuracy or ε-optimality for the unregularized (resp. regularized) case mean $V^{\star}(s) - V^{\pi^{(t)}}(s) \leq \varepsilon$ (resp. $V_{\tau}^{\star}(s) - V_{\tau}^{\pi^{(t)}}(s) \leq \varepsilon$) holds simultaneously for all $s \in \mathcal{S}$; ρ denotes the initial state distribution, which clearly obeys $\frac{1}{\min_{s\in\mathcal{S}} \rho(s)} \geq |\mathcal{S}|$. 5

1.2 Comparisons of last-iterate convergence of the proposed entropy-regularized PU and OMWU methods with prior results for finding ε-QRE or ε-NE of competitive matrix games. We note that the convergence rates of unregularized OMWU established in Wei et al. [2021a] are problem-dependent, and scale at least polynomially on the size of the action spaces. Desirable features in the last two columns are highlighted in blue. 10

1.3 Comparison of policy optimization methods for finding an ε-optimal NE (resp. QRE) of two-player zero-sum discounted Markov games in terms of the duality gap, i.e., a policy pair (μ, ν) satisfying $\max_{\mu',\nu'}(V^{\mu',\nu}(\rho) - V^{\mu,\nu'}(\rho)) \leq \varepsilon$ (resp. $\max_{\mu',\nu'}(V_{\tau}^{\mu',\nu}(\rho) - V_{\tau}^{\mu,\nu'}(\rho)) \leq \varepsilon$). Note that $*$ implies one-sided convergence, i.e., only one of the agents is guaranteed to achieve finite-time convergence to the equilibrium. Here, $c > 0$ refers to some problem-dependent constant. For simplicity and a fair comparison, we replace various notions of concentrability coefficient and distribution mismatch coefficient with a crude upper bound $\|1/\rho\|_{\infty}$, where ρ is the initial state distribution. 13

1.4 Comparison of policy optimization methods for finding an ε-optimal NE or QRE of two-player zero-sum episodic Markov games in terms of the duality gap. 14

1.5 Iteration complexities of the proposed OMWU method for finding ε-QRE/NE of zero-sum polymatrix games, where logarithmic dependencies are omitted. Here, γ denotes the maximal time delay when the delay is bounded, n denotes the number of agents in the game, d_{max} is the maximal degree of the graph, and $\|A\|_{\infty} = \max_{i,j} \|A_{i,j}\|_{\infty}$ is the ℓ_{∞} norm of the entire payoff matrix A (over all games in the network). We only present the result under statistical delay when the delays are bounded for ease of comparison, while more general bounds are given in Section 6.2.2. 16

List of Figures

1.1 Comparisons of PG and NPG methods with entropy regularization for a bandit problem ($\gamma = 0$) with 3 actions, whose corresponding rewards are 1.0, 0.9 and 0.1, respectively. The regularization parameter is set as $\tau = 0.1$ for the first row and $\tau = 1$ for the second row. In (a) and (d), the policy paths of $(\log \pi(a_1), \log \pi(a_2))$ following the PG method are plotted in orange, with the blue lines indicating the gradient flow; in (b) and (e), the policy paths of $(\log \pi(a_1), \log \pi(a_2))$ following the NPG method are depicted in red, with the blue lines indicating the natural gradient flow. The error contractions of both PG and NPG methods with $\eta = 0.1$ are shown in (c) and (f). 6

4.1 Performance illustration of the PU and OMWU methods for solving entropy-regularized matrix games with $|\mathcal{A}| = |\mathcal{B}| = 100$, where the entries of the payoff matrix A is generated independently from the uniform distribution on $[-1, 1]$. The learning rates are fixed as $\eta = 0.1$. The left panel plots various error metrics of convergence w.r.t. the iteration count with the entropy regularization parameter $\tau = 0.01$, while the right panel plots these error metrics at 1000-th iteration with different choices of τ. Due to their similar nature, PU and OMWU yield almost identical convergence behaviors and overlapping plots. 51

8.1 The cumulative regret v.s. number of iterations plot (left panel) and sub-optimality gap v.s. number of data plot (right panel) of VPO and MLE-MAB methods in the online and offline settings, respectively. 83

List of Algorithms

1 Entropy-regularized NPG with exact policy evaluation 29

2 PMD with generalized Bregman divergence (GPMD) 39
3 Approximate PMD with generalized Bregman divergence (Approximate GPMD) . . . 41

4 The PU method 49
5 The OMWU method 49

6 Entropy-regularized OMWU for Discounted Two-player Zero-sum Markov Game . . . 57
7 Entropy-regularized OMWU for Episodic Two-player Zero-sum Markov Game 59

8 Entropy-regularized OMWU, agent i 63

9 Independent NPG for Entropy-regularized Potential Games 72

10 VPO for online RLHF 80
11 VPO for offline RLHF 81

Chapter 1

Introduction

Reinforment learning (RL) is a machine learning paradigm focusing on sequential decision making problems where an agent learns to achieve a goal by interacting with its environment. At its essence, an RL agent learns to take the best action under any state input, so as to maximizes a numerical reward feedback signal. In contrast to the supervised learning paradigm, the optimal policy (state-action mapping) is typically not known to the system and hence needs to be identified by *exploring* the unknown environment.

RL has achieved tremendous success in a wide spectrum of applications, including strategic games [Mnih et al., 2015, Silver et al., 2016], robotic control [Lillicrap et al., 2016], AI alignment [Ouyang et al., 2022], to name a few. In contemporary RL applications, it is increasingly common to encounter environments with prohibitively large state and action spaces [Silver et al., 2016], which heightens the challenge of achieving efficient RL. This is particularly true in situations with limited data access due to high costs, time constraints, or high stakes, such as clinical trials [Liu et al., 2019b] and autonomous systems [Kiumarsi et al., 2017]. Furthermore, the intrinsic nonconcavity issues of RL formulations pose significant barriers to understanding the computational feasibility of current RL algorithms. These challenges naturally lead to two fundamental questions as researchers set out to establish the algorithmic foundations of RL:

- **Q1:** With an (estimated) environment model, how can we design practical algorithms to learn the optimal policy efficiently in terms of *iteration complexity*?
- **Q2:** How can we incorporate efficient exploring mechanism into the learning process, to improve the *sample complexity* of the learning system?

Regarding **Q1**, most of the recent empirical successes of RL can be attributed to the use of policy gradient (PG) methods and their variants [Williams, 1992, Sutton et al., 2000, Kakade, 2001, Peters and Schaal, 2008, Konda and Tsitsiklis, 2000]. In its basic form, the optimal policy of interest, or a suitably parameterized version, is learned by attempting to maximize the value function in a Markov decision processes (MDP), or achieving equilibrium in the presence of multiple agents. For the most part, the maximization step is carried out by means of first-order optimization algorithms amenable to large-scale applications, whose foundations were set forth in the early works of Williams [1992], Sutton et al. [2000]. A partial list of widely adopted variants in modern practice includes policy gradient (PG) methods [Sutton et al., 2000], natural policy gradient (NPG) methods [Kakade, 2001], TRPO [Schulman et al., 2015], PPO [Schulman et al., 2017b], soft actor-critic methods [Haarnoja et al., 2018], to name just a few. In comparison with model-based and value-based approaches, this family of policy-based algorithms offers a remarkably flexible framework that accommodates both

continuous and discrete action spaces, and lends itself well to the incorporation of powerful function approximation schemes like neural networks. In stark contrast to its practical success, however, theoretical understanding of policy optimization remains severely limited even for the tabular case, largely owing to the ubiquitous nonconvexity issue underlying the objective function, as well as the non-stationary nature of the multi-agent systems. Consequently, understanding and improving the computational efficiency of RL algorithms — sometimes coupled with additional resource and system-level constraints — inevitably lie at the core of cutting-edge RL research and are the key enabler for future advances.

The contemporary theory of reinforcement learning offers a conceptual roadmap towards addressing **Q2** by providing algorithm design principles that achieve (near) optimal sample complexity [Auer et al., 2008, Azar et al., 2017, Jin et al., 2018, Bai et al., 2019, Jiang et al., 2017]. These principles guide exploration towards less visited areas by constructing confidence sets of the underlying model and iteratively executing the policy derived from the most promising one. Despite strong theoretical guarantees, applying the optimism principle beyond the tabular case can be computationally intractable and thus not readily applicable to many real-world RL applications. On the other hand, many empirical heuristics promoting exploration lack theoretical validation. This has led to growing interest in developing principled exploratory algorithms that allow for efficient implementation and compatibility with deep learning architectures.

Throughout this thesis, we shall focus on:

- Designing efficient policy optimization methods with provable non-asymptotic *iteration complexity* for single-agent RL;
- Developing novel independent and symmetrical learning algorithms with improved *iteration complexity* for various multi-agent systems;
- Building efficient learning algorithms with provable non-asymptotic *sample complexity* for AI alignment task.

The rest of this chapter is organized as follows. Section 1.1 to Section 1.3 provide an overview of the main results of this thesis. Section 1.4 summarizes the related works. Finally, Section 1.5 provides the organization of the rest of the thesis.

1.1 Efficient policy optimization for single-agent RL

The goal of policy optimization in the single-agent RL setup is to maximize the value function that measures long-term cumulative reward. Despite the enormous empirical success, the theoretical underpinnings of policy gradient type methods have been limited even until recently, primarily due to the intrinsic non-concavity underlying the value maximization problem of interest [Bhandari and Russo, 2024, Agarwal et al., 2020b]. To further exacerbate the situation, an abundance of problem instances contain suboptimal policies residing in regions with flat curvatures (namely, vanishingly small gradients and high-order derivatives) [Agarwal et al., 2020b]. Such plateaus in the optimization landscape could, in principle, be difficult to escape once entered, thereby necessitating a higher degree of exploration in order to accelerate policy optimization.

1.1.1 Entropy-regularized RL

In practice, a strategy that has been frequently adopted to encourage exploration and improve convergence is to enforce entropy regularization [Williams and Peng, 1991, Peters et al., 2010,

Mnih et al., 2016, Duan et al., 2016, Haarnoja et al., 2017, Hazan et al., 2019, Vieillard et al., 2020, Xiao et al., 2019]. By inserting an additional penalty term to the objective function, this strategy penalizes policies that are not stochastic/exploratory enough, in the hope of preventing a policy optimization algorithm from being trapped in an undesired local region. Through empirical visualization, Ahmed et al. [2019] suggested that entropy regularization induces a smoother landscape that allows for the use of larger learning rates, and hence, faster convergence. However, the theoretical support for regularization-based policy optimization remains highly inadequate.

Motivated by this, a recent line of works set out to elucidate, in a theoretically sound manner, the efficiency of entropy-regularized policy gradient methods. Assuming access to exact policy gradients, Agarwal et al. [2020b] and Mei et al. [2020b] developed convergence guarantees for regularized PG methods (with relative entropy regularization considered in Agarwal et al. [2020b] and entropy regularization in Mei et al. [2020b]). Encouragingly, both papers suggested the positive role of regularization in guaranteeing faster convergence for the tabular setting. However, these works fell short of explaining the role of entropy regularization for other policy optimization algorithms like NPG methods, which we seek to understand in this thesis.

As an important and widely used extension of PG methods, *natural policy gradient* (NPG) methods propose to employ natural policy gradients [Amari, 1998] as search directions, in order to achieve faster convergence than the update rules based on policy gradients [Kakade, 2001, Peters and Schaal, 2008, Bhatnagar et al., 2009, Even-Dar et al., 2009]. Informally speaking, NPG methods precondition the gradient directions by Fisher information matrices (which are the Hessians of a certain divergence metric), and fall under the category of quasi second-order policy optimization methods. In fact, a variety of mainstream RL algorithms, such as *trust region policy optimization* (TRPO) [Schulman et al., 2015] and *proximal policy optimization* (PPO) [Schulman et al., 2017b], can be viewed as generalizations of NPG methods [Shani et al., 2020]. In this thesis, we pursue in-depth theoretical understanding about this popular class of methods — in conjunction with entropy regularization to be introduced momentarily.

Main contributions

Inspired by recent theoretical progress towards understanding PG methods [Agarwal et al., 2020b, Bhandari and Russo, 2024, Mei et al., 2020b], we aim to develop non-asymptotic convergence guarantees for entropy-regularized NPG methods in conjunction with softmax parameterization. We focus attention on studying tabular discounted Markov decision processes (MDPs), which is an important first step and a stepping stone towards demystifying the effectiveness of entropy-regularized policy optimization in more complex settings.

Settings. Consider a γ-discounted infinite-horizon MDP with state space $\mathcal{S}$ and action space $\mathcal{A}$. Assuming availability of exact policy evaluation, the update rule of entropy-regularized NPG methods with softmax parameterization admits a simple update rule in the policy space (see Section 2.1 for precise descriptions)

$$\pi^{(t+1)}(a|s) \;\propto\; \big(\pi^{(t)}(a|s)\big)^{1-\frac{\eta\tau}{1-\gamma}} \exp\Big(\frac{\eta Q_\tau^{\pi^{(t)}}(s,a)}{1-\gamma}\Big) \tag{1.1}$$

for any $(s,a) \in \mathcal{S}\times\mathcal{A}$, where $\tau > 0$ is the regularization parameter, $0 < \eta \le \frac{1-\gamma}{\tau}$ is the learning rate (or stepsize), $\pi^{(t)}$ indicates the t-th policy iterate, and Q_τ^π is the soft Q-function under policy π (to be defined in (2.10a)). The update rule (1.1) is closely connected to several popular algorithms in practice. For instance, the *trust region policy optimization* (TRPO) algorithm [Schulman et al.,

2015], when instantiated in the tabular setting, can be viewed as implementing (1.1) with line search. In addition, by setting the learning rate as $\eta = \frac{1-\gamma}{\tau}$, the update rule (1.1) coincides with *soft policy iteration* (SPI) studied in Haarnoja et al. [2017].

The results of this thesis deliver fully non-asymptotic convergence rates of entropy-regularized NPG methods without any hidden constants, which are previewed as follows (in an orderwise manner). The definition of ε-optimality can be found in Table 1.1.

- **Linear convergence of exact entropy-regularized NPG methods.** We establish linear convergence of entropy-regularized NPG methods for finding the optimal policy of the entropy-regularized MDP, assuming access to exact policy evaluation. To yield an ε-optimal policy for the regularized MDP (cf. Table 1.1), the algorithm (1.1) with a general learning rate $0 < \eta \leq \frac{1-\gamma}{\tau}$ needs no more than an order of

$$\frac{1}{\eta\tau} \log\left(\frac{1}{\varepsilon}\right)$$

  iterations, where we hide the dependencies that are logarithmic on salient problem parameters (see Theorem 1). Some highlights of our convergence results are (i) their near dimension-free feature and (ii) their applicability to a wide range of learning rates (including small learning rates).

- **Linear convergence of approximate entropy-regularized NPG methods.** We demonstrate the stability of the regularized NPG method with a general learning rate $0 < \eta \leq \frac{1-\gamma}{\tau}$ even when the soft Q-functions of interest are only available approximately. This paves the way for future investigations that involve finite-sample analysis. Informally speaking, the algorithm exhibits the same convergence behavior as in the exact gradient case before an error floor is hit, where the error floor scales linearly in the entrywise error of the soft Q-function estimates (see Theorem 2).

Comparisons with prior art. Agarwal et al. [2020b] proved that unregularized NPG methods with softmax parameterization attain an ε-accuracy within $\mathcal{O}(1/\varepsilon)$ iterations. In contrast, our results assert that $\mathcal{O}(\log(1/\varepsilon))$ iterations suffice with the assistance of entropy regularization, which hints at the potential benefit of entropy regularization in accelerating the convergence of NPG methods. Shortly after the initial posting of our paper, Bhandari and Russo [2021] posted a note that proves linear convergence of unregularized NPG methods with exact line search, by exploiting a clever connection to policy iteration. Their convergence rate is governed by a quantity $\min_{s\in\mathcal{S}} \rho(s)$, resulting in an iteration complexity at least $|\mathcal{S}|$ times larger than ours. In comparison, our results cover a broad range of fixed learning rates (including small stepsizes that are of particular interest in practice), and accommodate the scenario with inexact gradient evaluation. See Table 1.1 for a quantitative comparison. Moreover, we note that the entropy-regularized NPG method with general learning rates is closely related to TRPO in the tabular setting (see Shani et al. [2020]). The recent work Shani et al. [2020] demonstrated that TRPO converges with an iteration complexity $\mathcal{O}(1/\varepsilon)$ in entropy-regularized MDPs. The analysis therein is inspired by the mirror descent theory in generic optimization literature, which characterizes sublinear convergence under properly decaying stepsizes and accommodates various choices of divergence metrics. In comparison, our analysis strengthens the performance guarantees by carefully exploiting properties specific to the current version of the NPG method. In particular, we identify the delicate interplay between the crucial operational quantities $Q_\tau^\star - Q_\tau^{(t)}$ and $Q_\tau^\star - \tau \log \xi^{(t)}$ (to be defined later), and invoke the linear system theory to

Paper	Iteration complexity	Regularization	Learning rates
Agarwal et al. [2020b]	$\frac{2}{(1-\gamma)^2\varepsilon}+\frac{2}{\eta\varepsilon}$	unregularized	constant: $(0,\infty)$
Bhandari and Russo [2021]	$\frac{1}{(1-\gamma)\min_{s\in\mathcal{S}}\rho(s)}\log\left(\frac{1}{\varepsilon}\right)$	unregularized	exact line search
This Thesis	$\frac{1}{1-\gamma}\log\left(\frac{1}{\varepsilon}\right)$	regularized	constant: $\frac{1-\gamma}{\tau}$
This Thesis	$\frac{1}{\eta\tau}\log\left(\frac{1}{\varepsilon}\right)$	regularized	constant: $\left(0,\frac{1-\gamma}{\tau}\right)$

Table 1.1: The iteration complexities of NPG methods to reach ε-accuracy *in terms of optimization error*, where the unregularized (resp. regularized) version is given by (2.12) (cf. (2.14)) with η the learning rate. We assume exact gradient evaluation and softmax parameterization, and hide the dependencies that are logarithmic on problem parameters. Here, ε-accuracy or ε-optimality for the unregularized (resp. regularized) case mean $V^\star(s)-V^{\pi^{(t)}}(s)\le\varepsilon$ (resp. $V_\tau^\star(s)-V_\tau^{\pi^{(t)}}(s)\le\varepsilon$) holds simultaneously for all $s\in\mathcal{S}$; ρ denotes the initial state distribution, which clearly obeys $\frac{1}{\min_{s\in\mathcal{S}}\rho(s)}\ge|\mathcal{S}|$.

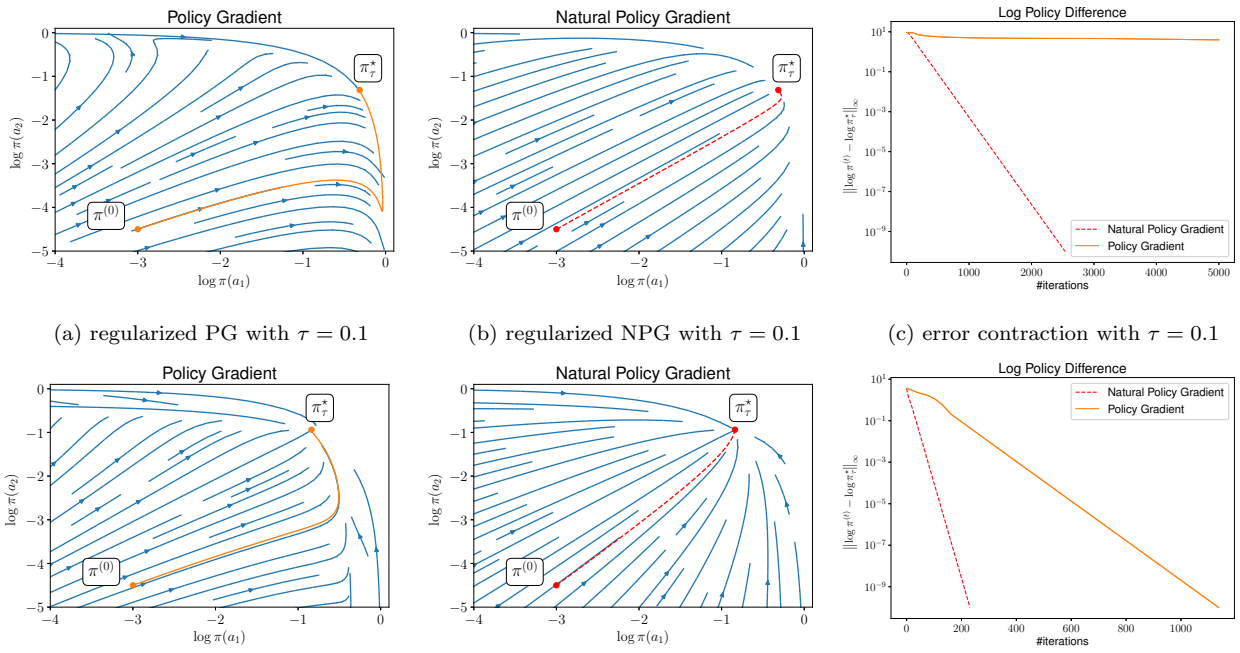
establish appealing contraction, which allow for the use of more aggressive constant stepsizes and hence improved convergence.

It is also helpful to compare our results with the state-of-the-art theory for PG methods with softmax parameterization [Agarwal et al., 2020b, Mei et al., 2020b]. Specifically, Agarwal et al. [2020b] established the asymptotic convergence of unregularized PG methods with softmax parameterization, while an iteration complexity of $\mathcal{O}(1/\varepsilon)$ was recently pinned down by Mei et al. [2020b]. In the presence of entropy regularization, Agarwal et al. [2020b] showed that PG with relative entropy regularization and softmax parameterization enjoys an iteration complexity of $\mathcal{O}(1/\varepsilon^2)$, while Mei et al. [2020b] showed that the entropy-regularized softmax PG method converges linearly in $\mathcal{O}(\log(1/\varepsilon))$ iterations. However, the dependencies of the iteration complexity in Mei et al. [2020b] on other salient parameters like $|\mathcal{S}|$, $|\mathcal{A}|$ and $\frac{1}{1-\gamma}$ are not fully specified. Very recently, Li et al. [2023] delivered a negative message demonstrating that these dependencies can be highly pessimistic; in fact, one can find an MDP instance which takes softmax PG methods (super)-exponential time (in terms of $|\mathcal{S}|$ and $\frac{1}{1-\gamma}$) to converge. In contrast, the bounds derived in the current paper are fully non-asymptotic, delineating clear dependencies on all salient problem parameters, which clearly demonstrate the algorithmic advantages of NPG methods. Fig. 1.1 depicts the policy paths of PG and NPG methods with entropy regularization for a simple bandit problem with three actions. It is evident from the plots that the NPG method follows a more direct path to the global optimum compared to the PG counterpart and hence converges faster. In addition, both algorithms converge more rapidly as the regularization parameter τ increases.

1.1.2 General regularized RL

In practice, there are often competing objectives and additional constraints that the agent has to deal with in conjunction with maximizing values, which motivate the studies of more general choices of regularization techniques in RL. In what follows, we isolate a few representative examples.

- *Promoting exploration.* In the face of large problem dimensions and complex dynamics, it is often desirable to maintain a suitable degree of randomness in the policy iterates, in order to encourage exploration and discourage premature convergence to sub-optimal policies. A popular strategy of this kind is to enforce entropy regularization [Williams and Peng, 1991], which penalizes policies that are not sufficiently stochastic. Along similar lines, the Tsallis



(a) regularized PG with $\tau = 0.1$ (b) regularized NPG with $\tau = 0.1$ (c) error contraction with $\tau = 0.1$

(d) regularized PG with $\tau = 1$ (e) regularized NPG with $\tau = 1$ (f) error contraction with $\tau = 1$

Figure 1.1: Comparisons of PG and NPG methods with entropy regularization for a bandit problem ($\gamma = 0$) with 3 actions, whose corresponding rewards are 1.0, 0.9 and 0.1, respectively. The regularization parameter is set as $\tau = 0.1$ for the first row and $\tau = 1$ for the second row. In (a) and (d), the policy paths of $(\log \pi(a_1), \log \pi(a_2))$ following the PG method are plotted in orange, with the blue lines indicating the gradient flow; in (b) and (e), the policy paths of $(\log \pi(a_1), \log \pi(a_2))$ following the NPG method are depicted in red, with the blue lines indicating the natural gradient flow. The error contractions of both PG and NPG methods with $\eta = 0.1$ are shown in (c) and (f).

entropy regularization [Chow et al., 2018b, Lee et al., 2018] further promotes sparsity of the learned policy while encouraging exploration, ensuring that the resulting policy does not assign non-negligible probabilities to too many sub-optimal actions.

- *Safe RL.* In a variety of application scenarios such as industrial robot arms and self-driving vehicles, the agents are required to operate safely both to themselves and the surroundings [Amodei et al., 2016, Moldovan and Abbeel, 2012]; for example, certain actions might be strictly forbidden in some states. One way to incorporate such prescribed operational constraints is through adding a regularizer (e.g., a properly chosen log barrier or indicator function tailored to the constraints) to explicitly account for the constraints.

- *Cost-sensitive RL.* In reality, different actions of an agent might incur drastically different costs even for the same state. This motivates the design of new objective functions that properly trade off the cumulative rewards against the accumulated cost, which often take the form of certain regularized value functions.

Viewed in this light, it is of imminent value to develop a unified framework towards understanding the capability and limitations of regularized policy optimization. While a recent line of works [Agarwal et al., 2020b, Mei et al., 2020b, Cen et al., 2022b] have looked into specific types of

regularization techniques such as entropy regularization, existing convergence theory remains highly inadequate when it comes to a more general family of regularizers.

Main contributions

We focus on policy optimization for regularized RL in a γ-discounted infinite horizon Markov decision process (MDP) with state space $\mathcal{S}$, action space $\mathcal{A}$, and reward function $r(\cdot,\cdot)$. The goal is to find an optimal policy that maximizes a regularized value function. Informally speaking, the regularized value function associated with a given policy π takes the following form:

$$V_\tau^\pi = V^\pi - \tau\mathbb{E}\big[h_s\big(\pi(\cdot\,|\,s)\big)\big],$$

where V^π denotes the original (unregularized) value function, $\tau > 0$ is the regularization parameter, $h_s(\cdot)$ denotes a convex regularizer employed to regularize the policy in state s, and the expectation is taken over certain marginal state distribution w.r.t. the MDP (to be made precise in Section 3.1.1). It is noteworthy that this thesis does not require the regularizer h_s to be either strongly convex or smooth.

In order to maximize the regularized value function (3.3b), Lan [2023] exhibited a seminal algorithm called *Policy Mirror Descent (PMD)*, which can be viewed as an adaptation of the mirror descent algorithm [Nemirovsky and Yudin, 1983, Beck and Teboulle, 2003] to the realm of policy optimization. In particular, PMD subsumes the natural policy gradient (NPG) method [Kakade, 2001] as a special case. To further generalize PMD [Lan, 2023], we propose an algorithm called *Generalized Policy Mirror Descent (GPMD)*. In each iteration, the policy is updated for each state in parallel via a mirror-descent style update rule. In sharp contrast to Lan [2023] that considered a generic Bregman divergence, our algorithm selects the Bregman divergence *adaptively* in cognizant of the regularizer, which leads to complementary perspectives and insights. Several important features and theoretical appeal of GPMD are summarized as follows.

- GPMD substantially broadens the range of (provably effective) algorithmic choices for regularized RL, and subsumes several well-known algorithms as special cases. For example, it reduces to regularized policy iteration [Geist et al., 2019] when the learning rate tends to infinity, and subsumes entropy-regularized NPG methods as special cases if we take the Bregman divergence to be the Kullback-Leibler (KL) divergence [Cen et al., 2022b].

- Assuming exact policy evaluation and perfect policy update in each iteration, GPMD converges linearly—in a dimension-free fashion— over the *entire* range of the learning rate $\eta > 0$. More precisely, it converges to an ε-optimal regularized Q-function in no more than an order of

  $$\frac{1+\eta\tau}{\eta\tau(1-\gamma)}\log\frac{1}{\varepsilon}$$

  iterations (up to some logarithmic factor). Encouragingly, this appealing feature is valid for a broad family of convex and possibly nonsmooth regularizers.

- The intriguing convergence guarantees are robust in the face of inexact policy evaluation and imperfect policy updates, namely, the algorithm is guaranteed to converge linearly at the same rate until an error floor is hit. See Section 3.2.2 for details.

- Numerical experiments demonstrate the practical applicability and appealing performance of the proposed GPMD algorithm.

Finally, we find it helpful to briefly compare the above findings with prior works. As soon as the learning rate exceeds $\eta \geq 1/\tau$, the iteration complexity of our algorithm is at most on the order of $\frac{1}{1-\gamma}\log\frac{1}{\varepsilon}$, thus matching that of regularized policy iteration [Geist et al., 2019]. In comparison to Lan [2023], our work sets forth a different framework to analyze mirror-descent type algorithms for regularized policy optimization, generalizing and refining the approach in Cen et al. [2022b] far beyond entropy regularization. When constant learning rates are employed, the linear convergence of PMD [Lan, 2023] critically requires the regularizer to be strongly convex, with only sublinear convergence theory established for convex regularizers. In contrast, we establish the linear convergence of GPMD under constant learning rates even in the absence of strong convexity. Furthermore, for the special case of entropy regularization, the stability analysis of GPMD also significantly improves over the prior art in Cen et al. [2022b], preventing the error floor from blowing up when the learning rate approaches zero, as well as incorporating the impact of optimization error that was previously uncaptured. More detailed comparisons with Lan [2023] and Cen et al. [2022b] can be found in Section 3.2.

1.2 Efficient policy optimization for multi-agent systems

Finding equilibria of multi-player games via gradient play lies at the heart of game theory, which permeates a remarkable breadth of modern applications, including but not limited to competitive reinforcement learning (RL) [Littman, 1994], generative adversarial networks (GANs) [Goodfellow et al., 2020] and adversarial training [Mertikopoulos et al., 2018b]. While it seems appealing to apply single-agent RL methods to each agent in a multi-agent system in a straightforward fashion, this approach neglects non-stationarity of the environment due to the presence of other agents, and thus lacks theoretical support in general. The complication has given rise to the paradigm of *centralized training with decentralized execution* (CTDE) [Lowe et al., 2017], where the policies are first obtained through training with a centralized controller with access to all agents' observations and then disseminated to each agent for execution. However, this approach falls short of adapting to changes in the environment without retraining and raises privacy concerns as well. It is hence of great interest to understand and design more versatile *independent learning* algorithms that only depend on the agents' local observations, require minimal coordination between agents, and provably converge.

1.2.1 Two-player zero-sum matrix games

We start by studying one of the most basic forms of multi-agent games, namely two-player zero-sum matrix games. Our goal is to find the equilibrium policies of both players in an *independent* and *decentralized* manner [Daskalakis et al., 2020, Wei et al., 2021b] with guaranteed *last-iterate convergence*. Namely, each player will execute symmetric and independent updates iteratively using its own payoff without observing the opponent's actions directly, and the final policies of the iterative process should be a close approximation to the equilibrium up to any prescribed precision. This kind of algorithms is more advantageous and versatile especially in federated environments, as it requires neither prior coordination between the players like two-timescale algorithms, nor a central controller to collect and disseminate the policies of all the players, which are often unavailable due to privacy constraints.

Last-iterate convergence in competitive games

In recent years, there have been significant progresses in understanding the last-iterate convergence of simple iterative algorithms for *unconstrained* saddle-point optimization, where one is interested in bounding the sub-optimality of the last iterate of the algorithm, rather than say, the ergodic iterate — which is the average of all the iterations — that are commonly studied in the earlier literature. This shift of focus is motivated, for example, by the infeasibility of averaging large machine learning models in training GANs [Goodfellow et al., 2020]. While vanilla Gradient Descent / Ascent (GDA) may diverge or cycle even for bilinear matrix games [Daskalakis et al., 2018], quite remarkably, small modifications lead to guaranteed last-iterate convergence to the equilibrium in a non-asymptotic fashion. A flurry of algorithms is proposed, including Optimistic Gradient Descent Ascent (OGDA) [Rakhlin and Sridharan, 2013, Daskalakis and Panageas, 2018, Wei et al., 2021a], predictive updates [Yadav et al., 2018], implicit updates [Liang and Stokes, 2019], and more. Several unified analyses of these algorithms have been carried out (see, e.g. Mokhtari et al. [2020a], Liang and Stokes [2019] and references therein), where these methods in principle all make clever extrapolation of the local curvature in a predictive manner to accelerate convergence. With slight abuse of terminology, in this thesis, we refer to this ensemble of algorithms as extragradient methods [Korpelevich, 1976, Tseng, 1995, Mertikopoulos et al., 2018a, Harker and Pang, 1990].

However, saddle-point optimization in the *constrained setting*, which includes competitive games as a special case, remains largely under-explored even for bilinear matrix games. While it is possible to reformulate constrained bilinear games to unconstrained ones using softmax parameterization of the probability simplex, this approach falls short of preserving the bilinear structure and convex-concave properties in the original problem, which are crucial to the convergence of gradient methods. Therefore, there is a strong necessity of understanding and developing improved extragradient methods in the constrained setting, where existing analyses in the unconstrained setting do not generalize straightforwardly. Daskalakis and Panageas [2019] proposed the optimistic variant of the multiplicative weight updates (MWU) method [Arora et al., 2012]—which is extremely natural and popular for optimizing over probability simplexes—called Optimistic Multiplicative Weight Updates (OMWU), and established the asymptotic last-iterate convergence of OMWU for matrix games. Very recently, Wei et al. [2021a] established non-asymptotic last-iterate convergences of OMWU. However, these last-iterate convergence results require the Nash equilibrium to be unique, and cannot be applied to problems with multiple Nash equilibria.

Main contributions

Motivated by the algorithmic role of entropy regularization in single-agent RL [Neu et al., 2017, Geist et al., 2019, Cen et al., 2022b], federated RL [Yang et al., 2023] as well as its wide use in game theory to account for imperfect and noisy information [McKelvey and Palfrey, 1995, Savas et al., 2019], we initiate the design and analysis of extragradient algorithms using *multiplicative updates* for finding the so-called quantal response equilibrium (QRE), which are solutions to competitive games with entropy regularization [McKelvey and Palfrey, 1995]. While finding QRE is of interest in its own right, by controlling the knob of entropy regularization, the QRE provides a close approximation to the Nash equilibrium (NE), and in turn acts as a smoothing scheme for finding the NE. Our contributions are summarized below, with the detailed problem formulations provided in Section 4.1.

- *Near dimension-free last-iterate convergence to QRE of entropy-regularized matrix games.* We propose two policy extragradient algorithms to solve entropy-regularized matrix games,

Equilibrium type	Method	Convergence rate	Dimension-free	Require unique NE
ε-QRE	PU & OMWU **This Thesis**	linear	yes	n/a
ε-NE	OMWU Daskalakis and Panageas [2019]	asymptotic	no	yes
	OMWU Wei et al. [2021a]	sublinear + linear	no	yes
	PU & OMWU **This Thesis**	sublinear	yes	no

Table 1.2: Comparisons of last-iterate convergence of the proposed entropy-regularized PU and OMWU methods with prior results for finding ε-QRE or ε-NE of competitive matrix games. We note that the convergence rates of unregularized OMWU established in Wei et al. [2021a] are problem-dependent, and scale at least polynomially on the size of the action spaces. Desirable features in the last two columns are highlighted in blue.

namely the Predictive Update (PU) and OMWU methods, where both players execute symmetric and multiplicative updates without knowing the entire payoff matrix nor the opponent's actions. Encouragingly, we show that the last iterate of the proposed algorithms converges to the unique QRE at a linear rate that is almost independent of the size of the action spaces. Roughly speaking, to find an ε-optimal QRE in terms of Kullback-Leibler (KL) divergence, it takes no more than

$$\widetilde{\mathcal{O}}\left(\frac{1}{\eta\tau}\log\left(\frac{1}{\varepsilon}\right)\right)$$

iterations, where $\widetilde{\mathcal{O}}(\cdot)$ hides logarithmic dependencies. Here, τ is the regularization parameter, and η is the learning rate of both players no larger than $\mathcal{O}(1/(\tau + \|A\|_\infty))$, where $\|A\|_\infty = \max_{i,j}|A_{i,j}|$ is the ℓ_∞ norm of the payoff matrix A. Optimizing the learning rate, the iteration complexity is bounded by $\widetilde{\mathcal{O}}\left(\|A\|_\infty\tau^{-1}\log(1/\varepsilon)\right)$.

- *Last-iterate convergence to ε-NE of unregularized matrix games without uniqueness assumption.* The QRE provides an accurate approximation to the NE by setting the entropy regularization τ sufficiently small, therefore our result directly translates to finding a NE with last-iterate convergence guarantee. Roughly speaking, to find an ε-NE (measured in terms of the duality gap), it takes no more than

$$\widetilde{\mathcal{O}}\left(\frac{\|A\|_\infty}{\varepsilon}\right)$$

iterations with optimized learning rates, again independent of the size of the action spaces up to logarithmic factors. Unlike prior literature [Daskalakis and Panageas, 2019, Wei et al., 2021a], our last-iterate convergence guarantee does not require the NE to be unique.

- *Extensions to two-player zero-sum Markov games.* By connecting value iteration with matrix games, we propose a policy extragradient method for solving infinite-horizon discounted

entropy-regularized zero-sum Markov games, which finds an ε-optimal minimax soft Q-function — in terms of ℓ_∞ error — in at most $\widetilde{\mathcal{O}}\left(\frac{1}{\tau(1-\gamma)^2}\log^2\left(\frac{1}{\varepsilon}\right)\right)$ iterations, where $\gamma \in (0,1)$ is the discount factor. By setting τ sufficiently small, the proposed method finds an ε-approximate NE (measured in terms of the duality gap) of the unregularized Markov game within $\widetilde{\mathcal{O}}\left(\frac{1}{(1-\gamma)^3\varepsilon}\right)$ iterations, which is independent of the dimension of the state-action space up to logarithmic factors.

To the best of our knowledge, this work is the first one that develops policy extragradient algorithms for solving entropy-regularized competitive games with multiplicative updates and dimension-free linear last-iterate convergence, and demonstrates entropy regularization as a smoothing technique to find ε-NE without the uniqueness assumption. Table 1.2 provides detailed comparisons of the proposed methods with prior arts for solving competitive games. Our results highlight the positive role of entropy regularization for accelerating convergence and safeguarding against imperfect payoff information in competitive games.

1.2.2 Two-player zero-sum Markov games

Two-player zero-sum Markov games [Shapley, 1953] generalizes two-player zero-sum matrix games by incoporating state transition dynamics and hence enabling modeling more complicated real-world problems. Substantial algorithmic developments have been made for finding equilibria in two-player zero-sum Markov games, where Dynamical Programming (DP) techniques have long been used as a fundamental building block, leading to prototypical iterative schemes such as Value Iteration (VI) [Shapley, 1953] and Policy Iteration (PI) [Van Der Wal, 1978, Patek and Bertsekas, 1999]. Different from their single-agent counterparts, these methods require solving a two-player zero-sum matrix game for every state per iteration. A considerable number of recent works [Zhao et al., 2022, Alacaoglu et al., 2022, Cen et al., 2021, Chen et al., 2021] are based on these DP iterations, by plugging in various (gradient-based) solvers of two-player zero-sum matrix games. However, these methods are inherently nested-loop, which barriers straightforward implementation. In addition, PI-based methods are asymmetric and come with only one-sided convergence guarantees [Patek and Bertsekas, 1999, Zhao et al., 2022, Alacaoglu et al., 2022].

This motivates us to design policy optimization algorithms that are *single-loop*, *symmetric*, with *finite-time last-iterate* convergence to the Nash Equilibrium (NE) or Quantal Response Equilibrium (QRE) under bounded rationality, two prevalent solution concepts in game theory. These design principles naturally come up as a result of pursuing simple yet efficient algorithms: *single-loop* updates preclude sophisticated interleaving of rounds between agents; *symmetric* updates ensure no agent will compromise its rewards in the learning process, which can be otherwise exploited by a faster-updating opponent; in addition, asymmetric updates typically lead to one-sided convergence, i.e., only one of the agents is guaranteed to converge to the minimax equilibrium in a non-asymptotic manner, which is less desirable; moreover, *last-iterate convergence* guarantee absolves the need for agents to switch between learning and deployment; last but not least, it is desirable to converge as fast as possible, where the iteration complexities are *non-asymptotic* with clear dependence on salient problem parameters.

Going beyond nested-loop algorithms, single-loop policy gradient methods have been proposed recently for solving two-player zero-sum Markov games. Here, we are interested in finding an ε-optimal NE or QRE in terms of the duality gap, i.e. the difference in the value functions when either of the agents deviates from the solution policy.

- For the infinite-horizon discounted setting, Daskalakis et al. [2020] demonstrated that the independent policy gradient method, with direct parameterization and asymmetric learning

rates, finds an ε-optimal NE within a polynomial number of iterations. Zeng et al. [2022] improved over this rate using an entropy-regularized policy gradient method with softmax parameterization and asymmetric learning rates. On the other end, Wei et al. [2021b] proposed an optimistic gradient descent ascent (OGDA) method [Rakhlin and Sridharan, 2013] with direct parameterization and symmetric learning rates,[1] which achieves a last-iterate convergence at a rather pessimistic iteration complexity.

- For the finite-horizon episodic setting, Zhang et al. [2022a], Yang and Ma [2023] showed that the weighted average-iterate of the optimistic Follow-The-Regularized-Leader (FTRL) method, when combined with slow critic updates, finds an ε-optimal NE in a polynomial number of iterations.

A more complete summary of prior results can be found in Table 1.3 and Table 1.4. In brief, while there have been encouraging progresses in developing computationally efficient policy gradient methods for solving zero-sum Markov games, achieving fast finite-time last-iterate convergence with single-loop and symmetric update rules remains a challenging goal.

Main contributions

Motivated by the positive role of entropy regularization in enabling faster convergence of policy optimization in single-agent RL [Cen et al., 2022b, Lan, 2023] and two-player zero-sum games [Cen et al., 2021], we propose a single-loop policy optimization algorithm for two-player zero-sum Markov games in both the infinite-horizon and finite-horizon settings. The proposed algorithm follows the style of actor-critic [Konda and Tsitsiklis, 2000], with the actor updating the policy via the entropy-regularized optimistic multiplicative weights update (OMWU) method [Cen et al., 2021] and the critic updating the value function on a slower timescale. Both agents execute multiplicative and symmetric policy updates, where the learning rates are carefully selected to ensure a fast last-iterate convergence. In both the infinite-horizon and finite-horizon settings, we prove that the last iterate of the proposed method learns the optimal value function and converges at a linear rate to the unique QRE of the entropy-regularized Markov game, which can be further translated into finding the NE by setting the regularization sufficiently small.

- For the infinite-horizon discounted setting, the last iterate of our method takes at most

$$\widetilde{\mathcal{O}}\left(\frac{|\mathcal{S}|}{(1-\gamma)^4\tau}\log\frac{1}{\varepsilon}\right)$$

iterations for finding an ε-optimal QRE under entropy regularization, where $\widetilde{\mathcal{O}}(\cdot)$ hides logarithmic dependencies. Here, $|\mathcal{S}|$ is the size of the state space, γ is the discount factor, and τ is the regularization parameter. Moreover, this implies the last-iterate convergence with an iteration complexity of

$$\widetilde{\mathcal{O}}\left(\frac{|\mathcal{S}|}{(1-\gamma)^5\varepsilon}\right)$$

for finding an ε-optimal NE.

[1] To be precise, Wei et al. [2021b] proved the average-iterate convergence of the duality gap, as well as the last-iterate convergence of the policy in terms of the Euclidean distance to the set of NEs, where it is possible to translate the latter last-iterate convergence to the duality gap. The resulting iteration complexity, however, is much worse than that of the average-iterate convergence in terms of the duality gap, with a problem-dependent constant that can scale pessimistically with salient problem parameters.

Solution type	Reference	Iteration complexity	Single loop	Symmetric	Last-iterate convergence
ε-NE	**PI-based Methods** **Zhao et al. [2022]** **Alacaoglu et al. [2022]**	$\widetilde{\mathcal{O}}\left(\frac{\|1/\rho\|_\infty}{(1-\gamma)^3\varepsilon}\right)^*$	✗	✗	✓
	VI-based Methods **Cen et al. [2021]** **Chen et al. [2021]**	$\widetilde{\mathcal{O}}\left(\frac{1}{(1-\gamma)^3\varepsilon}\right)$	✗	✓	✓
	Daskalakis et al. [2020]	Polynomial*	✓	✗	✗
	Zeng et al. [2022]	$\widetilde{\mathcal{O}}\left(\frac{\|\mathcal{S}\|^2\|1/\rho\|_\infty^5}{(1-\gamma)^{14}c^4\varepsilon^3}\right)^*$	✓	✗	✓
	Wei et al. [2021b]	$\widetilde{\mathcal{O}}\left(\frac{\|\mathcal{S}\|^3}{(1-\gamma)^9\varepsilon^2}\right)$	✓	✓	✗
		$\widetilde{\mathcal{O}}\left(\frac{\|\mathcal{S}\|^5(\|\mathcal{A}\|+\|\mathcal{B}\|)^{1/2}}{(1-\gamma)^{16}c^4\varepsilon^2}\right)$	✓	✓	✓
	This Thesis	$\widetilde{\mathcal{O}}\left(\frac{\|\mathcal{S}\|}{(1-\gamma)^5\varepsilon}\right)$	✓	✓	✓
ε-QRE	**VI-based Methods** **Cen et al. [2021]**	$\widetilde{\mathcal{O}}\left(\frac{1}{(1-\gamma)^3}\log^2\frac{1}{\varepsilon}\right)$	✗	✓	✓
	Zeng et al. [2022]	$\widetilde{\mathcal{O}}\left(\frac{\|\mathcal{S}\|^2\|1/\rho\|_\infty^5}{(1-\gamma)^{11}c^4\tau^3}\log\frac{1}{\varepsilon}\right)^*$	✓	✗	✓
	This Thesis	$\widetilde{\mathcal{O}}\left(\frac{\|\mathcal{S}\|}{(1-\gamma)^4\tau}\log\frac{1}{\varepsilon}\right)$	✓	✓	✓

Table 1.3: Comparison of policy optimization methods for finding an ε-optimal NE (resp. QRE) of two-player zero-sum discounted Markov games in terms of the duality gap, i.e., a policy pair (μ,ν) satisfying $\max_{\mu',\nu'}(V^{\mu',\nu}(\rho)-V^{\mu,\nu'}(\rho))\le\varepsilon$ (resp. $\max_{\mu',\nu'}(V_\tau^{\mu',\nu}(\rho)-V_\tau^{\mu,\nu'}(\rho))\le\varepsilon$). Note that $*$ implies one-sided convergence, i.e., only one of the agents is guaranteed to achieve finite-time convergence to the equilibrium. Here, $c>0$ refers to some problem-dependent constant. For simplicity and a fair comparison, we replace various notions of concentrability coefficient and distribution mismatch coefficient with a crude upper bound $\|1/\rho\|_\infty$, where ρ is the initial state distribution.

- For the finite-horizon episodic setting, the last iterate of our method takes at most

$$\widetilde{\mathcal{O}}\left(\frac{H^2}{\tau}\log\frac{1}{\varepsilon}\right)$$

iterations for finding an ε-optimal QRE under entropy regularization, where H is the horizon length. Similarly, this implies the last-iterate convergence with an iteration complexity of

$$\widetilde{\mathcal{O}}\left(\frac{H^3}{\varepsilon}\right)$$

for finding an ε-optimal NE.

Detailed comparisons between the proposed method and prior arts are provided in Table 1.3 and Table 1.4. To the best of our knowledge, this thesis presents the first method that is simultaneously single-loop, symmetric, and achieves fast finite-time last-iterate convergence in terms of the duality gap in both infinite-horizon and finite-horizon settings. From a technical perspective, the infinite-horizon discounted setting is in particular challenging, where ours is the first single-loop algorithm

Solution type	Reference	Iteration complexity	Single loop	Symmetric	Last-iterate convergence
ε-NE	**OFTRL** **Zhang et al. [2022a]**	$\widetilde{\mathcal{O}}\big(\frac{H^{28/5}}{\varepsilon^{6/5}}\big)$	✓	✓	✗
	Modified OFTRL **Zhang et al. [2022a]**	$\widetilde{\mathcal{O}}\big(\frac{H^4}{\varepsilon}\big)$	✓	✓	✗
	OFTRL **Yang and Ma [2023]**	$\widetilde{\mathcal{O}}\big(\frac{H^5}{\varepsilon}\big)$	✓	✓	✗
	This Thesis	$\widetilde{\mathcal{O}}\big(\frac{H^3}{\varepsilon}\big)$	✓	✓	✓
ε-QRE	**This Thesis**	$\widetilde{\mathcal{O}}\big(\frac{H^2}{\tau}\log\frac{1}{\varepsilon}\big)$	✓	✓	✓

Table 1.4: Comparison of policy optimization methods for finding an ε-optimal NE or QRE of two-player zero-sum episodic Markov games in terms of the duality gap.

that guarantees an iteration complexity of $\widetilde{\mathcal{O}}(1/\varepsilon)$ for last-iterate convergence in terms of the duality gap, with clear and improved dependencies on other problem parameters in the meantime. In contrast, several existing works introduce additional problem-dependent constants [Daskalakis et al., 2020, Wei et al., 2021b, Zeng et al., 2022] in the iteration complexity, which can scale rather pessimistically—sometimes even exponentially—with problem dimensions [Li et al., 2023].

Our technical developments require novel ingredients that deviate from prior tools such as error propagation analysis for Bellman operators [Perolat et al., 2015, Patek and Bertsekas, 1999] from a dynamic programming perspective, as well as the gradient dominance condition [Daskalakis et al., 2020, Zeng et al., 2022] from a policy optimization perspective. Importantly, at the core of our analysis lies a carefully-designed one-step error contraction bound for policy learning, together with a set of recursive error bounds for value learning, all of which tailored to the non-Euclidean OMWU update rules that have not been well studied in the setting of Markov games.

1.2.3 Multi-player zero-sum polymatrix games

In reality, there is no shortage of scenarios where the feedback can be obtained only in a delayed manner [He et al., 2014], i.e., the agents only receive the payoff information sent from a previous round instead of the current round, due to communication slowdowns and congestions, for example. Substantial progress has been made towards reliable and efficient online learning with delayed feedbacks in various settings, e.g., stochastic multi-armed bandit [Pike-Burke et al., 2018, Vernade et al., 2017], adversarial multi-armed bandit [Cesa-Bianchi et al., 2016, Li et al., 2019], online convex optimization [Quanrud and Khashabi, 2015, McMahan and Streeter, 2014] and multi-player game [Meng et al., 2023]. Typical approaches to combatting delays include subsampling the payoff history [Weinberger and Ordentlich, 2002, Joulani et al., 2013], or adopting adaptive learning rates suggested by delay-aware analysis [Quanrud and Khashabi, 2015, McMahan and Streeter, 2014, Hsieh et al., 2022, Flaspohler et al., 2021]. Most of these efforts, however, have been limited to the study of *individual regret*, which characterizes the performance gap between an agent's learning trajectory and the best policy in hindsight. It remains highly inadequate when it comes to guaranteeing *convergence* to the equilibrium in a multi-player environment, especially in the presence of delayed feedbacks, thus leaving the scalability and resiliency of gradient play open to questions.

In this work, we initiate the study of asynchronous learning algorithms for an important class of games called zero-sum polymatrix games (also known as network matrix games [Bergman and Fokin, 1998]), which generalizes two-player zero-sum matrix games to the multiple-player setting and serves as an important stepping stone to more general multi-player general-sum games. Zero-sum polymatrix games are commonly used to describe situations in which agents' interactions are captured by an interaction graph and the entire system of games are closed so that the total payoffs keep invariant in the system. They find applications in an increasing number of important domains such as security games [Cai et al., 2016], graph transduction [Bernardi, 2021], and more.

In particular, we focus on *finite-time last-iterate* convergence to two prevalent solution concepts in game theory, namely Nash Equilibrium (NE) and Quantal Response Equilibrium (QRE) which considers bounded rationality [McKelvey and Palfrey, 1995]. Despite the seemingly simple formulation, few existing works have achieved this goal even in the synchronous setting, i.e., with instantaneous feedback. Leonardos et al. [2021] studied a continuous-time learning dynamics that converges to the QRE at a linear rate. Anagnostides et al. [2022b] demonstrated Optimistic Mirror Descent (OMD) [Rakhlin and Sridharan, 2013] enjoys finite-time last-iterate convergence to the NE, yet the analysis therein requires continuous gradient of the regularizer, which incurs computation overhead for solving a subproblem every iteration. In contrast, an appealing alternative is the entropy regularizer, which leads to closed-form multiplicative updates and is computationally more desirable, but remains poorly understood. In sum, designing efficient learning algorithms that provably converge to the game equilibria has been technically challenging, even in the synchronous setting.

Main contributions

In this work, we develop provably convergent algorithms—broadly dubbed as *asynchronous gradient play*—to find the QRE and NE of zero-sum polymatrix games in a decentralized and symmetric manner with delayed feedbacks. We propose an entropy-regularized Optimistic Multiplicative Weights Update (OMWU) method [Cen et al., 2021], where each player symmetrically updates their strategies without access to the payoff matrices and other players' strategies, and initiate a systematic investigation on the impacts of delays on its convergence under two schemes of learning rates schedule. Our main contributions are summarized as follows.

- *Finite-time last-iterate convergence of single-timescale OMWU.* We begin by showing that, in the synchronous setting, the single-timescale OMWU method—when the same learning rate is adopted for extrapolation and update—achieves last-iterate convergence to the QRE at a linear rate, which is independent of the number of agents as well as the size of action spaces (up to logarithmic factors). In addition, this implies a last-iterate convergence to an ε-approximate NE in $\widetilde{\mathcal{O}}(\varepsilon^{-1})$ iterations by adjusting the regularization parameter, where $\widetilde{\mathcal{O}}(\cdot)$ hides logarithmic dependencies. While the last-iterate linear convergence to QRE continues to hold in the asynchronous setting, as long as the delay sequence follows certain mild statistical assumptions, it converges at a slower rate due to a smaller tolerable range of learning rates, with the iteration complexity to find an ε-NE degenerating to $\widetilde{\mathcal{O}}(\varepsilon^{-2})$. In addition, regret analysis of single-timescale OMWU is also provided.

- *Finite-time convergence of two-timescale OMWU.* To accelerate the convergence rate in the presence of delayed feedback, we propose a two-timescale OMWU method which separates the learning rates of extrapolation and update in a delay-aware manner for applications with constant and known delays (e.g. from timestamp information). The learning rate separation is critical in bypassing the convergence slowdown encountered in the single-timescale case,

Learning rate	Type of delay	Iteration complexity	
		ε-QRE	ε-NE
single-timescale	none	$\tau^{-1} d_{\mathsf{max}} \|A\|_\infty \log \varepsilon^{-1}$	$d_{\mathsf{max}} \|A\|_\infty \varepsilon^{-1}$
	statistical	$\tau^{-2} d_{\mathsf{max}}^2 \|A\|_\infty^2 (\gamma+1)^2 \log \varepsilon^{-1}$	$d_{\mathsf{max}}^2 \|A\|_\infty^2 (\gamma+1)^2 \varepsilon^{-2}$
two-timescale	constant	$\tau^{-1} d_{\mathsf{max}} \|A\|_\infty (\gamma+1)^2 \log \varepsilon^{-1}$	$d_{\mathsf{max}} \|A\|_\infty (\gamma+1)^2 \varepsilon^{-1}$
	bounded	$\tau^{-2} n d_{\mathsf{max}}^3 \|A\|_\infty^3 (\gamma+1)^{5/2} \varepsilon^{-1}$	$n d_{\mathsf{max}}^3 \|A\|_\infty^3 (\gamma+1)^{5/2} \varepsilon^{-3}$

Table 1.5: Iteration complexities of the proposed OMWU method for finding ε-QRE/NE of zero-sum polymatrix games, where logarithmic dependencies are omitted. Here, γ denotes the maximal time delay when the delay is bounded, n denotes the number of agents in the game, d_{max} is the maximal degree of the graph, and $\|A\|_\infty = \max_{i,j} \|A_{i,j}\|_\infty$ is the ℓ_∞ norm of the entire payoff matrix A (over all games in the network). We only present the result under statistical delay when the delays are bounded for ease of comparison, while more general bounds are given in Section 6.2.2.

where we show that two-timescale OMWU achieves a *faster* last-iterate linear convergence to QRE in the presence of constant delays, with an improved $\widetilde{\mathcal{O}}(\varepsilon^{-1})$ iteration complexity to ε-NE that matches the rate without delay. We further tackle the more practical yet challenging setting where the feedback sequence is permutated by bounded delays—possibly in an adversarial manner—and demonstrate provable convergence to the equilibria in an average-iterate manner.

We summarize the iteration complexities of the proposed methods for finding ε-approximate solutions of QRE and NE in Table 1.5. To the best of our knowledge, this thesis presents the first algorithm design and analysis that focus on equilibrium finding in a multi-player game with delayed feedbacks. In contrast, most of existing works concerning individual regret in the synchronous/asynchronous settings typically yield average-iterate convergence guarantees (see e.g., Bailey [2021], Meng et al. [2023]) and fall short of characterizing the actual learning trajectory to the equilibrium.

1.2.4 Multi-player potential games

Moving beyond competitive games, we focus on potential games [Monderer and Shapley, 1996b], an important class of games that admit a potential function to capture the differences in each agent's utility function induced by unilateral deviations. In particular, the analysis established in this thesis is tailored to potential games in their most basic setting, i.e., static potential games, an important stepping stone to the more general Markov setting. Despite its simple formulation and decades-long research, however, the computational underpinnings of such problems are still far from mature, especially when it comes to finding the Nash equilibrium (NE) of potential games in a decentralized manner. While several recent works have made significant breakthroughs by achieving logarithmic regrets with independent learning dynamics [Daskalakis et al., 2021, Anagnostides et al., 2022a], these results only guarantee convergence to coarse correlated equilibrium or correlated equilibrium, which are arguably much weaker equilibrium concepts than NE and hence do not lead to an approximate NE solution.

Main contributions

We seek to find the quantal response equilibrium (QRE) [McKelvey and Palfrey, 1995], the prototypical extension of NE for games with bounded rationality [Selten, 1989], where each agent runs independent natural policy gradient (NPG) methods [Kakade, 2001] involving symmetric, decentralized, and multiplicative updates according to its own payoff. This amounts to solving a potential game with entropy regularization, whose algorithmic role has been studied in the setting of single-agent RL [Mei et al., 2020b, Cen et al., 2022b] as well as two-player zero-sum games [Cen et al., 2021], but yet to be explored in more general settings. Our contributions are summarized below.

- **Finite-time global convergence of independent entropy-regularized NPG methods.** We show that independent entropy-regularized NPG methods provably converge to the QRE of a potential game, and it takes no more than

$$\mathcal{O}\left(\frac{\min\{\sqrt{N}, \Phi_{\max}\}\Phi_{\max}}{\tau^2\varepsilon^2}\right)$$

  iterations to find an ε-optimal QRE (to be defined precisely). Here, N stands for the number of agents, $\tau > 0$ the entropy regularization parameter, and $\Phi_{\max} > 0$ the maximum value of the potential function.

- **Finite-time global convergence to ε-NE without isolation assumption.** By setting the entropy regularization parameter τ sufficiently small, the result translates to finding an approximate NE with non-asymptotic convergence guarantees, thereby obviating the additional assumption in prior literature [Fox et al., 2022, Palaiopanos et al., 2017, Zhang et al., 2022b] that requires the set of stationary policies to be isolated. Specifically, it takes no more than

$$\widetilde{\mathcal{O}}\left(\frac{\min\{\sqrt{N}, \Phi_{\max}\}\Phi_{\max}}{\varepsilon^4}\right)$$

  iterations to find an ε-NE for the unregularized potential game, where $\widetilde{\mathcal{O}}$ hides logarithmic dependencies.

These rates give the first set of iteration complexities—to the best of our knowledge—that are independent of the size of the action spaces, up to logarithmic factors. In addition, the iteration complexities exhibit a sublinear dependency with the number of agents, outperforming existing NE-finding algorithms whose complexities depend at least linearly with the number of agents. Even more appealingly, when interpreting our convergence rates for the important special case of identical-interest games with bounded payoffs [Monderer and Shapley, 1996a], they further become independent with the number of agents, leading to the *first* method that achieves a *dimension-free* convergence rate of $\widetilde{\mathcal{O}}\left(1/\varepsilon^4\right)$ to find an ε-NE.

1.3 Principled policy optimization for AI alignment

Fine-tuning large language models (LLMs) by *reinforcement learning from human feedback* (RLHF) [Ziegler et al., 2019] has been shown to significantly improve the helpfulness, truthfulness and controllability of LLMs, as illustrated by InstructGPT [Ouyang et al., 2022] and many follow-ups. Roughly speaking, there are two critical components of RLHF: (1) *reward modeling*, which maps

human preference rankings into a quantitative reward function that can guide policy improvement; and (2) *RL fine-tuning*, which seeks to adjust LLM output to align with human preferences by leveraging the learned reward function, i.e., increasing the probability of preferred answers and decreasing the probability of unfavored answers.

Evidently, the curation of preference data is instrumental in the performance of RLHF, which is commonly modeled as pairwise comparisons from a Bradley-Terry ranking model [Bradley and Terry, 1952]. In particular, given a query x, human annotators choose a preferred answer from two candidate answers y_1 and y_2 generated by an LLM. Despite the simple form, collecting large-scale and high-quality preference data can be expensive and time-consuming. Depending on the availability of preference data, two paradigms of RLHF are considered: (1) *offline* RLHF, where only a pre-collected preference dataset is available, possibly generated from a pre-trained LLM after supervised fine-tuning (SFT); and (2) *online* RLHF, where additional preference data can be collected adaptively to improve alignment. While initial work on RLHF focused on the offline setting, the online setting has also begun to receive considerable attention, as even a small amount of additional preference data has been shown to greatly boost performance.

There has been significant work on the theoretical underpinnings of RLHF that seeks to uncover algorithmic improvements. Notably, while the original RLHF pipeline decouples reward modeling from RL fine-tuning, direct preference optimization (DPO) [Rafailov et al., 2023] integrates these as a single step in the *offline* setting, leveraging a closed-form solution for the optimal policy in the RL fine-tuning phase. This has led to a welcome simplification of the RLHF pipeline, allowing direct optimization of the policy (i.e., the LLM) from preference data.

Nevertheless, significant challenges remain in RLHF, particularly concerning how to incorporate estimates of reward *uncertainty* in direct preference optimization when parameterizing policies with large-scale neural networks — such as LLMs — in a theoretically and practically effective manner. In standard reinforcement learning (RL), managing uncertainty when an agent interacts with an environment is a critical aspect in achieving near-optimal performance [Sutton and Barto, 2018], when using methods that range from policy-based [Schulman et al., 2017b, Xiao et al., 2021], value-based [Mnih et al., 2015, Kumar et al., 2020], and actor-critic methods [Mnih et al., 2016]. One dominant approach in the bandit setting, for example, is to construct confidence intervals of the reward estimates, then acting according to the upper and lower confidence bounds — following the principles of optimism and pessimism in the online and offline settings respectively [Lattimore and Szepesvári, 2020, Lai et al., 1985, Rashidinejad et al., 2022].

Despite the fact that uncertainty estimation is even more critical in RLHF, due to the coarse nature of preference data, effective implementations of theoretically justified optimistic and pessimistic principles have yet to be developed in the RLHF literature. For example, existing online preference alignment methods, such as Nash-MD [Munos et al., 2023] and OAIF [Guo et al., 2024], do not incorporate exploration; similarly, pessimism is also not implemented in offline preference alignment methods, such as DPO [Rafailov et al., 2023] and IPO [Azar et al., 2024]. A key reason for these omissions is that it is extremely difficult to construct confidence intervals for arbitrary neural networks [Gawlikowski et al., 2023], let alone LLMs. Since optimism for online exploration and pessimism for offline RL both require uncertainty estimation, and given the difficulty of conducting uncertainty estimation for large-scale neural networks, a natural and important question arises:

Can we implement the optimistic/pessimistic principles under uncertainty in a practically efficient manner for online/offline preference alignment in LLMs while retaining theoretical guarantees?

Main contributions

In this thesis, we provide affirmative answer to the question. Our major contributions are as follows.

(i) We propose value-incentivized preference optimization (VPO) for both online and offline RLHF, a unified algorithmic framework that *directly optimizes the LLM policy* with the optimistic/pessimistic principles under uncertainty. Avoiding explicit uncertainty estimation, VPO regularizes maximum likelihood estimation of the reward function toward (resp. against) responses that lead to the highest value in the online (resp. offline) setting, hence implementing optimism (resp. pessimism). Theoretical regret guarantees of VPO are developed for both online and offline RLHF, matching their corresponding rates in the standard RL literature with explicit uncertainty estimation.

(ii) In addition, VPO reveals the critical role of reward calibration, where the shift ambiguity of the reward model inherent in the Bradley-Terry model [Bradley and Terry, 1952] can be exploited to implement additional behavior regularization [Pal et al., 2024, Ethayarajh et al., 2024]. This allows VPO to provide a theoretical foundation for popular conservative offline RL methods (e.g., [Kumar et al., 2020]), as well as regularized RLHF methods (e.g., DPOP [Pal et al., 2024]).

(iii) VPO admits a practically-implementable form suitable for RLHF on LLMs, and more generally, deep-learning architectures. We conduct extensive experimental studies using `TL;DR` and `ARC-Challenge` tasks in online and offline settings with optimistic and pessimistic bias, respectively. The results demonstrate improved empirical performance.

1.4 Related works

1.4.1 Single-agent RL

There has been a flurry of recent activities in studying theoretical behaviors of policy optimization methods. For example, Fazel et al. [2018], Jansch-Porto et al. [2020], Tu and Recht [2019], Zhang et al. [2020b], Mohammadi et al. [2021] established the global convergence of policy optimization methods for a couple of control problems (see the survey in Hu et al. [2023] for a comprehensive review of the latest developments in this area); Bhandari and Russo [2024] identified structural properties that guarantee the global optimality of PG methods without parameterization; Karimi et al. [2019] studied the convergence of PG methods to an approximate first-order stationary point, and Zhang et al. [2020d] proposed a variant of PG methods that converges to locally optimal policies leveraging saddle-point escaping algorithms in nonconvex optimization. Beyond the tabular setting, the convergence of PG methods with function approximations has been studied in Agarwal et al. [2020b], Wang et al. [2020], Liu et al. [2019a]. In particular, Cai et al. [2020] developed an optimistic variant of NPG that incorporates linear function approximation. We do not elaborate on this line of works since our focus is on understanding the performance of entropy-regularized NPG in the tabular setting; we also do not elaborate on PG methods that involve sample-based estimates, since we primarily consider exact gradients or black-box gradient estimators.

Regarding entropy regularization, Neu et al. [2017], Geist et al. [2019] provided unified views of entropy-regularized MDPs from an optimization perspective by connecting them to algorithms such as mirror descent [Nemirovsky and Yudin, 1983] and dual averaging [Nesterov, 2009]. The soft policy iteration algorithm has been identified as a special case of entropy-regularized NPG, highlighting again the link between policy gradient methods and soft Q-learning [Schulman et al.,

2017a]. The asymptotic convergence of soft policy iteration was established in Haarnoja et al. [2017], which fell short of providing explicit convergence rate guarantees. Additionally, Grill et al. [2019] developed planning algorithms for entropy-regularized MDPs, and Mei et al. [2020b] showed that the sub-optimality gap of soft policy iteration is small if the policy improvement is small in consecutive iterations.

Global convergence of policy gradient methods. Recent years have witnessed a surge of activities towards understanding the global convergence properties of policy gradient methods and their variants for both continuous and discrete RL problems, examples including Fazel et al. [2018], Bhandari and Russo [2024], Agarwal et al. [2020b], Zhang et al. [2020b], Wang et al. [2020], Mei et al. [2020a], Bhandari and Russo [2021], Khodadadian et al. [2021], Liu et al. [2020b], Mei et al. [2020a], Agazzi and Lu [2020], Xu et al. [2020a], Wang et al. [2020], Cen et al. [2022b], Mei et al. [2021], Liu et al. [2019a], Wang et al. [2021], Zhang et al. [2021b,a, 2020a], Shani et al. [2020], among other things. Neu et al. [2017] provided the first interpretation of NPG methods as mirror descent [Nemirovsky and Yudin, 1983], thereby enabling the adaptation of techniques for analyzing mirror descent to the studies of NPG-type algorithms such as TRPO [Shani et al., 2020, Tomar et al., 2022]. It has been shown that the NPG method converges sub-linearly for unregularized MDPs with a fixed learning rate [Agarwal et al., 2020b], and converges linearly if the learning rate is set adaptively [Khodadadian et al., 2021], via exact line search [Bhandari and Russo, 2021], or following a geometrically increasing schedule [Xiao, 2022]. Noteworthily, Li et al. [2023] established a lower bound indicating that softmax PG methods can take an exponential time—in the size of the state space—to converge, while the convergence rates of NPG-type methods are almost independent of the problem dimension. In addition, another line of recent works [Abbasi-Yadkori et al., 2019, Hao et al., 2021, Lazic et al., 2021] established regret bounds for approximate NPG methods—termed as KL-regularized approximate policy iteration therein—for infinite-horizen undiscounted MDPs, which are beyond the scope of this thesis.

Regularization in RL. Regularization has been suggested to the RL literature either through the lens of optimization [Dai et al., 2018, Agarwal et al., 2020b], or through the lens of dynamic programming [Geist et al., 2019, Vieillard et al., 2020]. Our work is clearly an instance of the former type. Several recent results in the literature merit particular attention: Agarwal et al. [2020b] demonstrated sublinear convergence guarantees for PG methods in the presence of relative entropy regularization, Mei et al. [2020b] established linear convergence of entropy-regularized PG methods. Most of the existing literature focused on the entropy regularization or KL-type regularization, and the studies of general regularizers had been quite limited until the recent work Lan [2023]. The regularized MDP problems are also closely related to the studies of constrained MDPs, as both types of problems can be employed to model/promote constraint satisfaction in RL, as recently investigated in, e.g., Chow et al. [2018a], Efroni et al. [2020], Ding et al. [2021], Yu et al. [2019], Xu et al. [2020b]. Note, however, that it is difficult to directly compare our algorithm with these methods, due to drastically different formulations and settings.

1.4.2 Multi-agent systems

Independent learning in general-sum games. Considerable progress has been made towards understanding independent learning dynamics in general-sum games Daskalakis et al. [2021], Anagnostides et al. [2022a] and general-sum Markov games (also known as stochastic games) Song et al. [2022], Jin et al. [2023], Mao and Başar [2022] by establishing non-asymptotic convergence to correlated equilibrium and coarse correlated equilibrium. However, such successes do not directly extend

to potential games where NE is of interest. Specialized analysis for potential games is thus needed as finding approximate NE in a two-player game can be PPAD-hard even with full information Daskalakis [2013]. Notably, there have been attempts to establish asymptotic convergence with independent learning dynamics Marden et al. [2007, 2009], Young [2004] for weakly acyclic games Young [2020], which includes potential games as a special case.

Learning in two-player zero-sum matrix games. Freund and Schapire [1999] showed that many standard methods such as GDA and MWU have a converging average duality gap at the rate of $\mathcal{O}(1/\sqrt{T})$, which is improved to $\mathcal{O}(1/T)$ by considering optimistic variants of these methods, such as OGDA and OMWU [Rakhlin and Sridharan, 2013, Daskalakis et al., 2011, Syrgkanis et al., 2015]. However, the last-iterate convergence of these methods are less understood until recently [Daskalakis and Panageas, 2019, Wei et al., 2021a]. In particular, under the assumption that the NE is unique for the unregularized matrix game, Daskalakis and Panageas [2019] showed the asymptotic convergence of the last iterate of OMWU to the unique equilibrium, and Wei et al. [2021a] showed the last iterate of OMWU achieves a linear rate of convergence after an initial phase of sublinear convergence, however the rates therein can be highly pessimistic in terms of the problem dimension, while our rate for entropy-regularized OMWU is dimension-free up to logarithmic factors. Sokota et al. [2023], Pattathil et al. [2023] showed that optimistic update is not necessary for achieving linear last-iterate convergence in the presence of regularization, albeit with a more strict restriction on the step size. In terms of no-regret analysis, Rakhlin and Sridharan [2013] established a no-regret learning rate of $\mathcal{O}(\log T/T^{1/2})$ with an auxiliary mixing of a uniform distribution at each update, which is later improved to $\mathcal{O}(1/T^{1/2})$ in Kangarshahi et al. [2018] with a slightly different algorithm.

Learning in two-player zero-sum Markov games. In addition to the aforementioned works on policy optimization methods (policy-based methods) for two-player zero-sum Markov games (cf. Table 1.3 and Table 1.4), a growing body of works have developed model-based methods [Liu et al., 2021, Zhang et al., 2020c, Li et al., 2022] and value-based methods [Bai and Jin, 2020, Bai et al., 2020, Chen et al., 2022, Jin et al., 2023, Sayin et al., 2021, Xie et al., 2020], with a primary focus on learning NE in a sample-efficient manner. Our work, together with prior literatures on policy optimization, focuses instead on learning NE in a computation-efficient manner assuming full-information.

Saddle-point optimization. Considerable progress has been made towards understanding OGDA and extragradient (EG) methods in the unconstrained convex-concave saddle-point optimization with general objective functions [Mokhtari et al., 2020a,b, Nemirovski, 2004, Liang and Stokes, 2019]. However, most works have focused on either average-iterate convergence (also known as ergodic convergence) [Nemirovski, 2004], or the characterization of *Euclidean update* rules [Mokhtari et al., 2020a,b, Liang and Stokes, 2019], where parameters are updated in an additive manner. These analyses do not generalize in a straightforward manner to *non-Euclidean* updates. As a result, the last-iterate convergence of non-Euclidean updates for saddle-point optimization still lacks theoretical understanding in general, and most works fall short of characterizing a finite-time convergence result. In particular, Mertikopoulos et al. [2018a] demonstrated the asymptotic last-iterate convergence of EG, and Hsieh et al. [2019] investigated similar questions for single-call EG algorithms. Lei et al. [2021] showed that OMWU converges to the equilibrium locally without an explicit rate. Wei et al. [2021a] showed that the last-iterate of OGDA converges linearly for strongly-convex strongly-concave constrained saddle-point optimization with an explicit rate.

Entropy regularization in games. Entropy regularization has been used to account for imperfect information in the seminal work of McKelvey and Palfrey [1995] that introduced the QRE, and a few representative works on entropy and more general regularizations in games include but are not limited to Savas et al. [2019], Hofbauer and Sandholm [2002], Mertikopoulos and Sandholm [2016].

Policy optimization in potential games. Fox et al. [2022], Palaiopanos et al. [2017], Zhang et al. [2022b] established asymptotic convergence of independent NPG methods for Markov potential games with an additional assumption that requires the set of stationary policies to be isolated. Heliou et al. [2017] demonstrated asymptotic convergence of NPG with diminishing step sizes for potential games in the bandit feedback setting. In addition, Zhang et al. [2022b] proposed to use a log-barrier regularization along with NPG to sidestep the isolation assumption and achieved the same iteration complexity as that of PG methods with direct parameterization. In contrast, we consider NPG with entropy regularization, which achieves a convergence rate that has better dependencies on the size of the action spaces and the number of agents.

1.4.3 AI alignment

RLHF. Since the introduction of the original RLHF framework, there have been many proposed simplifications of the preference alignment procedure and attempts to improve performance, including but not limited to SLiC [Zhao et al., 2023], GSHF [Xiong et al., 2023], DPO [Rafailov et al., 2023], and its variants, such as Nash-MD [Munos et al., 2023], IPO [Azar et al., 2024], OAIF [Guo et al., 2024], SPO [Swamy et al., 2024], SPIN [Chen et al., 2024], GPO [Tang et al., 2024], and DPOP [Pal et al., 2024]. These methods can roughly be grouped into online and offline variants, depending on whether preference data is collected before training (offline) or by using the current policy during training (online).

In offline preference alignment, identity preference optimization (IPO, [Azar et al., 2024]) argues that it is problematic to use the Bradley-Terry model in DPO to convert pairwise preferences into pointwise reward values, and proposes an alternative objective function to bypass the use of the Bradley-Terry model. DPO-Positive (DPOP, [Pal et al., 2024]) observes a failure mode of DPO that the standard DPO loss can reduce the model's likelihood on preferred answers, and proposes to add a regularization term to the DPO objective to avoid such a failure mode. On the other hand, online AI feedback (OAIF, [Guo et al., 2024]) proposes an online version of DPO, where online preference data from LLM annotators is used to evaluate and update the current LLM policy in an iterative manner. Iterative reasoning preference optimization (Iterative RPO, Pang et al. [2024]) proposes to add an additional negative log-likelihood term in the DPO loss to improve performances on reasoning tasks. Finally, Chang et al. [2024] proposes to reuse the offline preference data via reset.

Reward-biased exploration in RL. Reward-biased maximum likelihood estimation (RBMLE) promotes exploration by incoporating a bias term associated with the optimal value into the likelihood function. Kumar and Becker [1982] initiated the study of the RBMLE principle and proved asymptotically convergence to optimal long-term reward for solving unknown MDPs. The approach has been shown to achieve order-optimal finite-time regret bounds in multi-armed bandit problems [Liu et al., 2020a, Hung et al., 2021] and online RL [Mete et al., 2021, Liu et al., 2024a] recently. VPO draws inspiration from reward-biased exploration in the standard online RL literature, but significantly broadens its scope to the offline setting and RLHF for the first time.

Concurrent work on principled RLHF. Since posting the initial version of this work on arXiv, we discovered several concurrent work that also appeared online around the same time proposing similar regularization techniques as ours to encourage optimism (resp. pessimism) for online (resp. offline) RLHF [Zhang et al., 2024b, Xie et al., 2024, Liu et al., 2024b]. In the context of online RLHF, [Zhang et al., 2024b] empirically studies the similar algorithm as the proposed online VPO under the contextual bandit formulation of RLHF; Xie et al. [2024] provides finite-time regret analysis of the similar algorithm for the token-level MDP formulation with general function approximation, which extends to general deterministic contextual MDP as well. In the context of offline RLHF, Liu et al. [2024b] studies the similar algorithm as the proposed offline VPO and provides sample complexity analysis under the same contextual bandit formulation, yet focuses on general function approximation and different assumptions.

1.5 Thesis organization and notation

The rest of this thesis prospectus is organized as follows. Part I focuses on the theoretical development of policy optimization for single-agent RL, where Chapter 2 and 3 provide the algorithms and theories for learning entropy-regularized RL and general regularized RL, respectively. Part II covers the main results on various multi-agent systems, with Chapter 4, 5, 6, 7 focusing on two-player zero-sum matrix game, two-player zero-sum Markov game, multi-player polymatix game and multi-player potential game, respectively.

We denote by $\Delta(\mathcal{S})$ (resp. $\Delta(\mathcal{A})$) the probability simplex over the set $\mathcal{S}$ (resp. $\mathcal{A}$). When scalar functions such as $|\cdot|$, $\exp(\cdot)$ and $\log(\cdot)$ are applied to vectors, their applications should be understood in an entry-wise fashion. For instance, given any vector $z = [z_i]_{1\le i\le n} \in \mathbb{R}^n$, the notation $|\cdot|$ denotes $|z| \coloneqq [|z_i|]_{1\le i\le n}$; other functions are defined analogously. For any vectors $z = [z_i]_{1\le i\le n}$ and $w = [w_i]_{1\le i\le n}$, the notation $z \ge w$ (resp. $z \le w$) means $z_i \ge w_i$ (resp. $z_i \le w_i$) for all $1 \le i \le n$. The softmax function $\mathsf{softmax} : \mathbb{R}^n \mapsto \mathbb{R}^n$ is defined such that $[\mathsf{softmax}(\theta)]_i \coloneqq \exp(\theta_i)/\big(\sum_i \exp(\theta_i)\big)$ for a vector $\theta = [\theta_i]_{1\le i\le n} \in \mathbb{R}^n$. For any convex and differentiable function $h(\cdot)$, the Bregman divergence generated by $h(\cdot)$ is defined as

$$D_h(z, x) \coloneqq h(z) - h(x) - \big\langle \nabla h(x), z - x \big\rangle. \tag{1.2}$$

For any convex (but not necessarily differentiable) function $h(\cdot)$, we denote by ∂h the subdifferential of h. Given two probability distributions π_1 and π_2 over $\mathcal{A}$, the Kullback-Leibler (KL) divergence from π_2 to π_1 is defined by $\mathsf{KL}\left(\pi_1 \,\|\, \pi_2\right) \coloneqq \sum_{a\in\mathcal{A}} \pi_1(a) \log \frac{\pi_1(a)}{\pi_2(a)}$. Given two probability distributions p and q over $\mathcal{S}$, we introduce the notation $\left\|\frac{p}{q}\right\|_\infty \coloneqq \max_{s\in\mathcal{S}} \frac{p(s)}{q(s)}$ and $\left\|\frac{1}{q}\right\|_\infty \coloneqq \max_{s\in\mathcal{S}} \frac{1}{q(s)}$. Given a matrix A, $\|A\|_\infty$ is used to denote entrywise maximum norm, namely, $\|A\|_\infty = \max_{i,j} |A_{i,j}|$. The all-one vector is denoted as $\mathbf{1}$. We denote Jeffrey divergence [Jeffreys, 1998] by $J(\pi, \pi') = \mathsf{KL}\left(\pi \,\|\, \pi'\right) + \mathsf{KL}\left(\pi' \,\|\, \pi\right)$, which is the symmetric version of the KL divergence. For a vector $\mathbf{a} \in \mathcal{A}^N$, we use $a_i \in \mathcal{A}$ and $a_{-i} \in \mathcal{A}^{N-1}$ to denote the entry with index i and all the rest entries as a vector, respectively.

Part I

Policy optimization for single-agent RL

Chapter 2

Entropy-regularized Natural Policy Gradient Method

In this section, we formulate the problem of policy optimization for single-agent RL, as well as the non-asymptotic convergence guarantee for entropy-regularized NPG method. For more details and entire analysis, please refer to Cen et al. [2022b].

2.1 Model and algorithms

2.1.1 Problem settings

Markov decision processes. We focus on a discounted Markov decision process (MDP) [Puterman, 2014] denoted by $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, \gamma)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $\gamma \in (0,1)$ indicates the discount factor, $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the transition kernel, and $r : \mathcal{S} \times \mathcal{A} \to [0,1]$ stands for the reward function.[1] To be more specific, for each state-action pair $(s,a) \in \mathcal{S} \times \mathcal{A}$ and any state $s' \in \mathcal{S}$, we denote by $P(s'|s,a)$ the transition probability from state s to state s' when action a is taken, and $r(s,a)$ the instantaneous reward received in state s due to action a. A policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ represents a (randomized) action selection rule, namely, $\pi(a|s)$ specifies the probability of executing action a in state s for each $(s,a) \in \mathcal{S} \times \mathcal{A}$.

Value functions and Q-functions. For any given policy π, we denote by $V^\pi : \mathcal{S} \to \mathbb{R}$ the corresponding value function, namely, the expected discounted cumulative reward with an initial state $s_0 = s$, given by

$$\forall s \in \mathcal{S}: \qquad V^\pi(s) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \,\middle|\, s_0 = s\right], \tag{2.1}$$

where the action $a_t \sim \pi(\cdot|s_t)$ follows the policy π and $s_{t+1} \sim P(\cdot|s_t, a_t)$ is generated by the MDP $\mathcal{M}$ for all $t \geq 0$. We also overload the notation $V^\pi(\rho)$ to indicate the expected value function of a policy π when the initial state is drawn from a distribution ρ over $\mathcal{S}$, namely,

$$V^\pi(\rho) := \mathbb{E}_{s\sim\rho}\left[V^\pi(s)\right]. \tag{2.2}$$

[1] For the sake of simplicity, we assume throughout that the reward resides within $[0,1]$. Our results can be generalized in a straightforward manner to other ranges of bounded rewards.

Additionally, the Q-function $Q^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ of a policy π — namely, the expected discounted cumulative reward with an initial state $s_0 = s$ and an initial action $a_0 = a$ — is defined by

$$\forall (s,a) \in \mathcal{S} \times \mathcal{A}: \qquad Q^\pi(s,a) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t,a_t) \,\big|\, s_0 = s, a_0 = a\right], \tag{2.3}$$

where the action $a_t \sim \pi(\cdot|s_t)$ follows the policy π for all $t \geq 1$, and $s_{t+1} \sim P(\cdot|s_t,a_t)$ is generated by the MDP $\mathcal{M}$ for all $t \geq 0$.

Discounted state visitation distributions. A type of marginal distributions — commonly dubbed as *discounted state visitation distributions* — plays an important role in our theoretical development. To be specific, the discounted state visitation distribution $d_{s_0}^\pi$ of a policy π given the initial state $s_0 \in \mathcal{S}$ is defined by

$$\forall s \in \mathcal{S}: \qquad d_{s_0}^\pi(s) := (1-\gamma)\sum_{t=0}^{\infty} \gamma^t \mathbb{P}(s_t = s \mid s_0), \tag{2.4}$$

where the trajectory $(s_0, s_1, \cdots)$ is generated by the MDP $\mathcal{M}$ under policy π starting from state s_0. In words, $d_{s_0}^\pi(\cdot)$ captures the state occupancy probabilities when each state visitation is properly discounted depending on the time stamp. Further, for any distribution ρ over $\mathcal{S}$, we define the distribution d_ρ^π as follows

$$\forall s \in \mathcal{S}: \qquad d_\rho^\pi(s) := \mathbb{E}_{s_0 \sim \rho}\big[d_{s_0}^\pi(s)\big], \tag{2.5}$$

which describes the discounted state visitation distribution when the initial state s_0 is randomly drawn from a prescribed initial distribution ρ.

Softmax parameterization. It is common practice to parameterize the class of feasible policies in a way that is amenable to policy optimization. The focal point of this thesis is softmax parameterization — a widely adopted scheme which naturally ensures that the policy lies in the probability simplex. Specifically, for any $\theta : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ (called "logic values"), the corresponding softmax policy π_θ is generated through the softmax transform

$$\pi_\theta := \mathsf{softmax}(\theta) \qquad \text{or} \qquad \forall (s,a) \in \mathcal{S} \times \mathcal{A}: \quad \pi_\theta(a|s) := \frac{\exp(\theta(s,a))}{\sum_{a' \in \mathcal{A}} \exp(\theta(s,a'))}. \tag{2.6}$$

In what follows, we shall often abuse the notation to treat π_θ and θ as vectors in $\mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$, and suppress the subscript θ from π_θ, whenever it is clear from the context.

Entropy-regularized value maximization. To promote exploration and discourage premature convergence to suboptimal policies, a widely used strategy is entropy regularization, which searches for a policy that maximizes the following entropy-regularized value function

$$V_\tau^\pi(\rho) := V^\pi(\rho) + \tau \cdot \mathcal{H}(\rho,\pi). \tag{2.7}$$

Here, the quantity $\tau \geq 0$ denotes the regularization parameter, and $\mathcal{H}(\rho,\pi)$ stands for a sort of *discounted entropy* defined as follows

$$\mathcal{H}(\rho,\pi) := \mathop{\mathbb{E}}_{\substack{s_0 \sim \rho, a_t \sim \pi(\cdot|s_t),\\ s_{t+1} \sim P(\cdot|s_t,a_t), \forall t \geq 0}} \left[\sum_{t=0}^{\infty} -\gamma^t \log \pi(a_t|s_t)\right] = \frac{1}{1-\gamma} \mathop{\mathbb{E}}_{s \sim d_\rho^\pi}\left[\sum_{a \in \mathcal{A}} \pi(a|s) \log \frac{1}{\pi(a|s)}\right]. \tag{2.8}$$

Equivalently, V_τ^π can be viewed as the value function of π by adjusting the instantaneous reward to be policy-dependent regularized version as follows

$$\forall (s,a) \in \mathcal{S} \times \mathcal{A}: \qquad r_\tau(s,a) := r(s,a) - \tau \log \pi(a|s). \tag{2.9}$$

We also define $V_\tau^\pi(s)$ analogously when the initial state is fixed to be any given state $s \in \mathcal{S}$. The regularized Q-function Q_τ^π of a policy π, also known as the soft Q-function,[2] is related to V_τ^π as

$$\forall (s,a) \in \mathcal{S} \times \mathcal{A}: \qquad Q_\tau^\pi(s,a) = r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} \big[V_\tau^\pi(s') \big], \tag{2.10a}$$

$$\forall s \in \mathcal{S}: \qquad V_\tau^\pi(s) = \mathbb{E}_{a \sim \pi(\cdot|s)} \big[-\tau \log \pi(a|s) + Q_\tau^\pi(s,a) \big]. \tag{2.10b}$$

Optimal policies and stationary distributions. Denote by $\pi^\star$ (resp. $\pi_\tau^\star$) the policy that maximizes the value function (resp. regularized value function with regularization parameter τ), and let $V^\star$ (resp. $V_\tau^\star$) represent the resulting optimal value function (resp. regularized value function). Importantly, the optimal policies $\pi^\star$ and $\pi_\tau^\star$ of the MDP do not depend on the initial distribution ρ [Mei et al., 2020b]. In addition, $\pi^\star$ and $\pi_\tau^\star$ maximize the Q-function and the soft Q-function, respectively (which is self-evident from (2.10a)). A simple yet crucial connection between $\pi^\star$ and $\pi_\tau^\star$ can be demonstrated via the following sandwich bound[3]

$$V^{\pi_\tau^\star}(\rho) \le V^{\pi_\star}(\rho) \le V^{\pi_\tau^\star}(\rho) + \frac{\tau}{1-\gamma} \log |\mathcal{A}|, \tag{2.11}$$

which holds for all initial distributions ρ. The key takeaway message is that: the optimal policy $\pi_\tau^\star$ of the regularized problem could also be nearly optimal in terms of the unregularized value function, as long as the regularization parameter τ is chosen to be sufficiently small.

2.1.2 Algorithm: NPG methods with entropy regularization

Natural policy gradient methods. Towards computing the optimal policy (in the parameterized form), perhaps the first strategy that comes into mind is to run gradient ascent w.r.t. the parameter θ until convergence — a first-order method commonly referred to as the *policy gradient* (PG) algorithm (e.g. Sutton et al. [2000]). In comparison, the *natural policy gradient* (NPG) method [Kakade, 2001] adopts a pre-conditioned gradient update rule

$$\theta \;\leftarrow\; \theta + \eta \big(\mathcal{F}_\rho^\theta\big)^\dagger \nabla_\theta V^{\pi_\theta}(\rho), \tag{2.12}$$

in the hope of searching along a direction independent of the policy parameterization in use. Here, η is the learning rate or stepsize, $\mathcal{F}_\rho^\theta$ denotes the Fisher information matrix given by

$$\mathcal{F}_\rho^\theta := \mathop{\mathbb{E}}_{s \sim d_\rho^{\pi_\theta}, a \sim \pi_\theta(\cdot|s)} \Big[\big(\nabla_\theta \log \pi_\theta(a|s)\big) \big(\nabla_\theta \log \pi_\theta(a|s)\big)^\top \Big], \tag{2.13}$$

and we use $B^\dagger$ to indicate the Moore-Penrose pseudoinverse of a matrix B. It has been understood that the NPG method essentially attempts to monitor/control the policy changes approximately in terms of the Kullback-Leibler (KL) divergence (see e.g. Schulman et al. [2015, Section 7]).

[2] In this thesis, we use the terms "regularized" value (resp. Q) functions and "soft" value (resp. Q) functions interchangeably.

[3] To see this, invoke the optimality of $\pi_\tau^\star$ and the elementary entropy bound $0 \le \mathcal{H}(\rho, \pi) \le \frac{1}{1-\gamma} \log |\mathcal{A}|$ to obtain

$$V^{\pi_\tau^\star}(\rho) + \tfrac{\tau}{1-\gamma} \log |\mathcal{A}| \ge V^{\pi_\tau^\star}(\rho) + \tau \mathcal{H}(\rho, \pi_\tau^\star) = V_\tau^\star(\rho) \ge V_\tau^{\pi_\star}(\rho) \ge V^{\pi_\star}(\rho).$$

NPG methods with entropy regularization. Equipped with entropy regularization, the NPG update rule can be written as

$$\theta \;\leftarrow\; \theta + \eta\big(\mathcal{F}_\rho^\theta\big)^\dagger \nabla_\theta V_\tau^{\pi_\theta}(\rho), \tag{2.14}$$

where $\mathcal{F}_\rho^\theta$ is defined in (2.13) and $V_\tau^\pi(\rho)$ is defined in (2.7). Under softmax parameterization, this update rule admits a fairly simple form in the policy space , which, interestingly, is invariant to the choice of ρ. More precisely, if we let $\theta^{(t)}$ denote the t-th iterate and $\pi^{(t)} = \mathsf{softmax}(\theta^{(t)})$ the associated policy, then the entropy-regularized NPG updates satisfy

$$\pi^{(t+1)}(a|s) \;=\; \frac{1}{Z^{(t)}(s)}\big(\pi^{(t)}(a|s)\big)^{1-\frac{\eta\tau}{1-\gamma}} \exp\Big(\frac{\eta Q_\tau^{\pi^{(t)}}(s,a)}{1-\gamma}\Big), \tag{2.15}$$

where $Q_\tau^{\pi^{(t)}}$ is the soft Q-function of policy $\pi^{(t)}$, and $Z^{(t)}(s)$ is some normalization factor. This can alternatively be viewed as an instantiation/variant of the *trust region policy optimization* (TRPO) algorithm (see Schulman et al. [2015], Shani et al. [2020]). As an important special case, the update rule (2.15) reduces to

$$\pi^{(t+1)}(\cdot|s) \;=\; \frac{1}{Z^{(t)}(s)} \exp\Big(\frac{Q_\tau^{\pi^{(t)}}(s,\cdot)}{\tau}\Big) \qquad \text{when } \eta = \frac{1-\gamma}{\tau} \tag{2.16}$$

for some normalization factor $Z^{(t)}(s)$. The procedure (2.16) can be interpreted as a "soft" version of the classical policy iteration algorithm [Bertsekas, 2017] (as it employs a softmax function to approximate the max operator) w.r.t. the soft Q-function, and is often dubbed as *soft policy iteration* (SPI) (see Haarnoja et al. [2018, Section 4.1]).

To simplify notation, we shall use $V_\tau^{(t)}$, $Q_\tau^{(t)}$ and $d_\rho^{(t)}$ throughout to denote $V_\tau^{\pi^{(t)}}$, $Q_\tau^{\pi^{(t)}}$ and $d_\rho^{\pi^{(t)}}$, respectively. The complete procedure is summarized in Algorithm 1.

Algorithm 1: Entropy-regularized NPG with exact policy evaluation

inputs: learning rate η, initialization $\pi^{(0)}$.
for $t = 0, 1, 2, \cdots$ **do**
  Compute the regularized Q-function $Q_\tau^{(t)}$ (defined in (2.10a)) of policy $\pi^{(t)}$.
  Update the policy:

$$\forall (s,a) \in \mathcal{S}\times\mathcal{A}: \quad \pi^{(t+1)}(a|s) = \frac{1}{Z^{(t)}(s)}\big(\pi^{(t)}(a|s)\big)^{1-\frac{\eta\tau}{1-\gamma}} \exp\Big(\frac{\eta Q_\tau^{(t)}(s,a)}{1-\gamma}\Big), \tag{2.17}$$

  where $Z^{(t)}(s) = \sum_{a'\in\mathcal{A}}\big(\pi^{(t)}(a'|s)\big)^{1-\frac{\eta\tau}{1-\gamma}} \exp\big(\frac{\eta Q_\tau^{(t)}(s,a')}{1-\gamma}\big)$.

2.2 Main results

2.2.1 Exact entropy-regularized NPG methods

We first study the convergence behavior of entropy-regularized NPG methods (2.17) assuming access to exact policy evaluation in every iteration (namely, we assume the soft Q-function $Q_\tau^{(t)}$ can be evaluated accurately in all t). Remarkably, this algorithm converges linearly — in terms of computing both the optimal soft Q-function $Q_\tau^\star$ and the associated log policy $\log \pi_\tau^\star$ — as asserted by the following theorem.

Theorem 1 (Linear convergence of exact entropy-regularized NPG)**.** *For any learning rate* $0 < \eta \le (1-\gamma)/\tau$*, the entropy-regularized NPG updates* (2.17) *satisfy*

$$\left\|Q_\tau^\star - Q_\tau^{(t+1)}\right\|_\infty \le C_1 \gamma \left(1-\eta\tau\right)^t \tag{2.18a}$$

$$\left\|\log \pi_\tau^\star - \log \pi^{(t+1)}\right\|_\infty \le 2C_1 \tau^{-1} (1-\eta\tau)^t \tag{2.18b}$$

for all $t \ge 0$*, where*

$$C_1 := \left\|Q_\tau^\star - Q_\tau^{(0)}\right\|_\infty + 2\tau\left(1 - \frac{\eta\tau}{1-\gamma}\right)\left\|\log \pi_\tau^\star - \log \pi^{(0)}\right\|_\infty. \tag{2.19}$$

It is worth emphasizing that Theorem 1 is stated in a completely non-asymptotic form containing *no* hidden constants, and that our result covers any learning rate η in the range $(0, (1-\gamma)/\tau]$. A few implications of this theorem are in order.

- **Linear convergence of soft Q-functions.** To reach $\left\|Q_\tau^\star - Q_\tau^{(t)}\right\|_\infty \le \varepsilon$, the entropy-regularized NPG method needs at most $\frac{1}{\eta\tau}\log\left(\frac{C_1\gamma}{\varepsilon}\right)$ iterations. Remarkably, the iteration complexity almost does not depend on the dimensions of the MDP (except for some very weak dependency embedded in $\log C_1$) — this inherits a dimension-free feature of NPG methods that has been highlighted in Agarwal et al. [2020b] for the unregularized case. When the learning rate η is fixed in the admissible range, the iteration complexity scales inverse proportionally with τ, suggesting a higher level of entropy regularization might accelerate convergence, albeit to the solution of a regularized problem that is further away from the original MDP.

- **Linear convergence of log policies.** In contrast to the unregularized case, entropy regularization ensures uniqueness of the optimal policy and, therefore, makes it possible to study the convergence of the policy directly. Our theorem reveals that the entropy-regularized NPG method needs at most $\frac{1}{\eta\tau}\log\left(\frac{2C_1}{\varepsilon\tau}\right)$ iterations to yield $\left\|\log \pi_\tau^\star - \log \pi^{(t+1)}\right\|_\infty \le \varepsilon$.

- **Linear convergence of soft value functions.** As a byproduct, Theorem 1 implies that the iterates of soft value functions also converge linearly, namely,

$$\left\|V_\tau^\star - V_\tau^{(t+1)}\right\|_\infty \le (\gamma+2)C_1\left(1-\eta\tau\right)^t. \tag{2.20}$$

  To see this, we make note of the following relation previously established in Nachum et al. [2017]:

$$\forall (s,a) \in \mathcal{S}\times\mathcal{A}: \qquad V_\tau^\star(s) = -\tau\log\pi_\tau^\star(a|s) + Q_\tau^\star(s,a),$$

$$\Longrightarrow \qquad V_\tau^\star(s) = \mathop{\mathbb{E}}_{a\sim\pi^{(t+1)}(\cdot|s)}\left[-\tau\log\pi_\tau^\star(a|s) + Q_\tau^\star(s,a)\right].$$

  Consequently, combining this with the definition (2.10b) yields

$$\begin{aligned}\left|V_\tau^\star(s) - V_\tau^{(t+1)}(s)\right| &= \mathop{\mathbb{E}}_{a\sim\pi^{(t+1)}(\cdot|s)}\left[\left(-\tau\log\pi_\tau^\star(a|s) + Q_\tau^\star(s,a)\right) - \left(-\tau\log\pi_\tau^{(t+1)}(a|s) + Q_\tau^{(t+1)}(s,a)\right)\right] \\ &\le \tau\left\|\log\pi_\tau^\star - \log\pi_\tau^{(t+1)}\right\|_\infty + \left\|Q_\tau^\star - Q_\tau^{(t+1)}\right\|_\infty,\end{aligned}$$

  which together with (2.18) immediately establishes (2.20).

- **Convergence rate of SPI.** The best convergence guarantee is achieved when $\eta = (1-\gamma)/\tau$ (i.e. the SPI case), where the iteration complexity to reach $\left\|Q_\tau^\star - Q_\tau^{(t)}\right\|_\infty \le \varepsilon$ reduces to

$$\frac{1}{1-\gamma}\log\left(\frac{\gamma\left\|Q_\tau^\star - Q_\tau^{(0)}\right\|_\infty}{\varepsilon}\right),$$

which is proportional to the effective horizon $\frac{1}{1-\gamma}$ modulo some log factor. This means the iteration complexity of SPI recovers that of policy iteration [Puterman, 2014]. Interestingly, the contraction rate in this case (which is γ) is independent of the choice of the regularization parameter τ. Similarly, the iteration complexity of SPI to reach $\left\|\log\pi_\tau^\star - \log\pi^{(t+1)}\right\|_\infty \le \varepsilon$ becomes $\frac{1}{1-\gamma}\log\left(\frac{2\|Q_\tau^\star - Q_\tau^{(0)}\|_\infty}{\varepsilon\tau}\right)$, and the contraction rate is again independent of τ.

Comparison with entropy-regularized policy gradient methods. Mei et al. [2020b, Theorem 6] proved that the entropy-regularized policy gradient method achieves[4]

$$V_\tau^\star(\rho) - V_\tau^{(t)}(\rho) \le \left(V_\tau^\star(\rho) - V_\tau^{(0)}(\rho)\right) \cdot \exp\left(-\frac{(1-\gamma)^4 t}{(8/\tau + 4 + 8\log|\mathcal{A}|)|\mathcal{S}|}\left\|\frac{d_\rho^{\pi_\tau^\star}}{\rho}\right\|_\infty^{-1} \min_s \rho(s)\left(\inf_{0\le k\le t-1}\min_{s,a}\pi^{(k)}(a|s)\right)^2\right),$$

and they further showed that $\inf_{k\ge 0}\min_{s,a}\pi^{(k)}(a|s)$ is non-vanishing in t. It remains unclear, however, how $\inf_{t\ge 0}\min_{s,a}\pi^{(t)}(a|s)$ scales with other potentially large salient parameters like $|\mathcal{S}|$, $|\mathcal{A}|$, $\frac{1}{1-\gamma}$, $\frac{1}{\tau}$. In truth, existing theory does not rule out the possibility of exponential dependency on these salient parameters. It would thus be of great interest to establish algorithm-dependent lower bounds to uncover the right scaling with these important parameters. In contrast, our convergence guarantees for entropy-regularized NPG methods unveil concrete dependencies on all problem parameters.

Computing an ε-optimal policy for the original MDP. Thus far, we have established an intriguing convergence behavior of the entropy-regularized NPG method. However, caution needs to be exercised when interpreting the efficacy of this method: the preceding results are concerned with convergence to the optimal regularized value function $V_\tau^\star$, as opposed to finding the optimal value function $V^\star$ of the original MDP. Fortunately, by choosing the regularization parameter τ to be sufficiently small (in accordance with the target accuracy level ε), we can guarantee that $V_\tau^\star \approx V^\star$ (cf. (2.11)), thus ensuring the relevance and applicability of our results for solving the original MDP. To be specific, let us adopt the following choice of τ:

$$\tau = \frac{(1-\gamma)\varepsilon}{4\log|\mathcal{A}|}, \tag{2.21}$$

and assume the error of the regularized value function satisfies $\left\|V_\tau^\star - V_\tau^{(t)}\right\|_\infty < \varepsilon/2$. By virtue of Theorem 1, this optimization accuracy can be achieved via no more than $\frac{4\log|\mathcal{A}|}{(1-\gamma)\eta\varepsilon}\log\left(\frac{2C_1\gamma}{\varepsilon}\right)$ iterations of entropy-regularized NPG updates with a general learning rate,[5] or no more than

[4] Here, we have assumed the exact policy gradient is computed with respect to $V_\tau^{(t)}(\rho)$.

[5] This result is in fact better than the iteration complexity $\frac{2}{(1-\gamma)^2\varepsilon}$ of the unregularized NPG method established in Agarwal et al. [2020b] as soon as $\eta \ge 2(1-\gamma)\log|\mathcal{A}|\log\left(\frac{2C_1\gamma}{\varepsilon}\right)$. Consequently, our finding hints at the potential advantage of entropy-regularized NPG methods over the unregularized counterpart even when solving the original MDP.

$\frac{1}{1-\gamma}\log\left(\frac{\gamma\left\|Q_\tau^\star - Q_\tau^{(0)}\right\|_\infty}{\varepsilon}\right)$ iterations with the specific choice $\eta = \frac{1-\gamma}{\tau}$. It then follows that

$$\begin{aligned} V^\star(s) - V^{(t)}(s) &= V^\star(s) - V_\tau^\star(s) + V_\tau^\star(s) - V_\tau^{(t)}(s) + V_\tau^{(t)}(s) - V^{(t)}(s) \\ &\le \left(V^\star(s) - V_\tau^\star(s)\right) + \left\|V_\tau^\star - V_\tau^{(t)}\right\|_\infty + \left(V_\tau^{(t)}(s) - V^{(t)}(s)\right) \\ &\le \frac{2\tau\log|\mathcal{A}|}{1-\gamma} + \frac{\varepsilon}{2} = \varepsilon \end{aligned}$$

for any $s \in \mathcal{S}$, where we have used our choice of τ in (2.21). Here, the second inequality arises from (2.11) as well as the fact that for any policy π,

$$\left\|V_\tau^\pi - V^\pi\right\|_\infty = \tau \max_s \left|\mathcal{H}(s,\pi)\right| \le \frac{\tau\log|\mathcal{A}|}{1-\gamma},$$

given the elementary entropy bound $0 \le \mathcal{H}(s,\pi) \le \frac{1}{1-\gamma}\log|\mathcal{A}|$.

2.2.2 Approximate entropy-regularized NPG methods

There is no shortage of scenarios where the soft Q-function $Q_\tau^{(t)}(s,a)$ is available only in an approximate fashion, e.g. the cases when the value function has to be evaluated using finite samples. To account for inexactness of policy evaluation, we extend our theory to accommodate the following approximate update rule: for any $s \in \mathcal{S}$ and any $t \ge 0$,

$$\pi^{(t+1)}(\cdot|s) \;\propto\; \left(\pi^{(t)}(\cdot|s)\right)^{1-\frac{\eta\tau}{1-\gamma}} \exp\left(\frac{\eta\widehat{Q}_\tau^{(t)}(s,\cdot)}{1-\gamma}\right), \quad \text{where} \quad \left\|\widehat{Q}_\tau^{(t)} - Q_\tau^{(t)}\right\|_\infty \le \delta. \tag{2.22}$$

Here, δ is some quantity that captures the size of approximation errors. We do not specify the estimator for the soft Q-function (as long as it satisfies the entrywise estimation bound), thus allowing one to plug in both model-based and model-free value function estimators designed for a variety of sampling mechanisms (e.g. Azar et al. [2013], Li et al. [2021]). Encouragingly, the algorithm (2.22) is robust vis-à-vis inexactness of value function estimates, as it still converges linearly until an error floor is hit.

Theorem 2 (Linear convergence of approximate entropy-regularized NPG)**.** *When $0 < \eta \le (1-\gamma)/\tau$, the inexact entropy-regularized NPG updates (2.22) satisfy*

$$\left\|Q_\tau^\star - Q_\tau^{(t+1)}\right\|_\infty \le \gamma\left[(1-\eta\tau)^t C_1 + C_2\right] \tag{2.23a}$$

$$\left\|\log\pi_\tau^\star - \log\pi^{(t+1)}\right\|_\infty \le 2\tau^{-1}\left[(1-\eta\tau)^t C_1 + C_2\right] \tag{2.23b}$$

for all $t \ge 0$, where C_1 is the same as defined in (2.19) and C_2 is given by

$$C_2 := \frac{2\delta}{1-\gamma}\left(1 + \frac{\gamma}{\eta\tau}\right) = \frac{2\delta}{(1-\gamma)^2}\left[1 + \gamma\left(\frac{1-\gamma}{\eta\tau} - 1\right)\right]. \tag{2.24}$$

Apparently, Theorem 2 reduces to Theorem 1 when $\delta = 0$. As implied by this theorem, if the ℓ_∞ error of the soft-Q function estimates does not exceed

$$\delta \le \frac{(1-\gamma)^2\varepsilon}{2\gamma\left[1 + \gamma\left(\frac{1-\gamma}{\eta\tau} - 1\right)\right]},$$

then the algorithm (2.22) achieves 2ε-accuracy (i.e. $\left\|Q_\tau^\star - Q_\tau^{(t)}\right\|_\infty \le 2\varepsilon$) within $\frac{1}{\eta\tau}\log\left(\frac{C_1\gamma}{\varepsilon}\right)$ iterations. In particular, in the case of soft policy iteration (i.e. $\eta = \frac{1-\gamma}{\tau}$), the tolerance level δ can be up to $\frac{(1-\gamma)^2\varepsilon}{2\gamma}$, which matches the theory of approximate policy iteration in Agarwal et al. [2019].

Remark 1. *It is straightforward to combine Theorem 2 with known sample complexities for approximate policy evaluation to obtain a crude sample complexity bound. For instance, assuming access to a generative model, Li et al. [2024b] asserts that for any fixed policy π, model-based policy evaluation achieves $\left\|\widehat{Q}_\tau^\pi - Q_\tau^\pi\right\|_\infty \le \delta$ with high probability, as long as the number of samples per state-action pair exceeds the order of*

$$\frac{1}{(1-\gamma)^3\delta^2}$$

up to some logarithmic factor. By employing fresh samples for each policy evaluation, we can set $\delta = \frac{(1-\gamma)^2\varepsilon}{2\gamma}$ and invoke the union bound over $\widetilde{\mathcal{O}}\big(\frac{1}{1-\gamma}\big)$ iterations to demonstrate that: SPI with model-based policy evaluation needs at most

$$\widetilde{\mathcal{O}}\left(\frac{|\mathcal{S}|\,|\mathcal{A}|}{(1-\gamma)^8\varepsilon^2}\right)$$

samples to find an ε-optimal policy. Here, $\widetilde{\mathcal{O}}(\cdot)$ hides any logarithmic factor. We note, however, that the above sample analysis is extremely crude and might be improvable by, say, allowing sample reuses across iterations. It remains an interesting open question as to whether NPG with entropy regularization is minimax-optimal with a generative model, where the minimax lower bound is on the order of $\frac{|\mathcal{S}|\,|\mathcal{A}|}{(1-\gamma)^3\varepsilon^2}$ [Azar et al., 2013] and achievable by model-based plug-in estimators [Agarwal et al., 2020a, Li et al., 2024b] but not by vanilla Q-learning [Li et al., 2024a].

2.3 Discussion

This thesis establishes non-asymptotic convergence of entropy-regularized natural policy gradient methods, providing theoretical footings for the role of entropy regularization in guaranteeing fast convergence. Our analysis opens up several directions for future research; we close the paper by sampling a few of them.

- *Extended analysis of policy gradient methods with inexact gradients.* It would be of interest to see whether our analysis framework can be applied to improve the theory of policy gradient methods [Mei et al., 2020b] to accommodate the case with inexact policy gradients.

- *Finite-sample analysis in the presence of sample-based policy evaluation.* Another natural extension is towards understanding the sample complexity of entropy-regularized NPG methods when the value functions are estimated using rollout trajectories (see e.g. Kakade and Langford [2002], Agarwal et al. [2020b], Shani et al. [2020]), or using bootstrapping (see e.g. Xu et al. [2020c], Haarnoja et al. [2018], Wu et al. [2020]).

- *Function approximation.* The current work has been limited to the tabular setting. It would certainly be interesting, and fundamentally important, to understand entropy-regularized NPG methods in conjunction with function approximation; see Sutton et al. [2000], Agarwal et al. [2019, 2020b] for a few representative scenarios.

- *Beyond softmax parameterization.* The current paper has been devoted to softmax parameterization, which enables a concise and NPG update rule. A couple of other parameterization schemes have been proposed for (vanilla) PG methods as well [Agarwal et al., 2019, 2020b, Bhandari and Russo, 2024, 2021], e.g. vanilla parameterization (paired with proper projection onto the probability simplex in each iteration), log-linear parameterization, and neural softmax parameterization. Unfortunately, the analysis in our paper relies heavily on the

softmax NPG update rule, and does not immediately extend to other parameterization. It would be of great importance to establish convergence guarantees that accommodate other parameterizations of practical interest.

Chapter 3

Generalized Policy Mirror Descent Method

In this section, we focus on the general framework of regularized RL that subsumes entropy regularization as an special example. We present the proposed general policy mirror descent (GPMD) method, as well as the accompanying theory confirming its convergence to the optimal policy. For more details and entire analysis, please refer to Zhan et al. [2023a].

3.1 Model and algorithms

3.1.1 Problem settings

Regularized MDP. In practice, the agent is often asked to design policies that possess certain structural properties in order to be cognizant of system constraints such as safety and operational constraints, as well as encourage exploration during the optimization/learning stage. A natural strategy to achieve these is to resort to the following *regularized value function* w.r.t. a given policy π [Neu et al., 2017, Mei et al., 2020b, Cen et al., 2022b, Lan, 2023]:

$$\forall s \in \mathcal{S}: \qquad V_\tau^\pi(s) \coloneqq \mathop{\mathbb{E}}_{\substack{a_t\sim\pi(\cdot|s_t),\\ s_{t+1}\sim P(\cdot|s_t,a_t),\ \forall t\geq 0}} \left[\sum_{t=0}^{\infty}\gamma^t\Big\{r(s_t,a_t)-\tau h_{s_t}\big(\pi(\cdot\,|\,s_t)\big)\Big\}\ \Big|\ s_0=s\right]$$
$$= V^\pi(s) - \frac{\tau}{1-\gamma}\sum_{s'\in\mathcal{S}} d_s^\pi(s')h_{s'}\big(\pi(\cdot\,|\,s')\big), \tag{3.1}$$

where $h_s:\Delta_\zeta(\mathcal{A})\to\mathbb{R}$ stands for a convex and possibly nonsmooth regularizer for state s, $\tau>0$ denotes the regularization parameter, and $d_s^\pi(\cdot)$ is defined in (2.4). Here, for technical convenience, we assume throughout that $h_s(\cdot)$ $(s\in\mathcal{S})$ is well-defined over an "ζ-neighborhood" of the probability simplex $\Delta(\mathcal{A})$ defined as follows

$$\Delta_\zeta(\mathcal{A}) \coloneqq \left\{x=[x_a]_{a\in\mathcal{A}}\ \Big|\ x_a\geq 0 \text{ for all } a\in\mathcal{A};\ 1-\zeta\leq\sum_{a\in\mathcal{A}}x_a\leq 1+\zeta\right\},$$

where $\zeta>0$ can be an arbitrary constant. For instance, entropy regularization adopts the choice $h_s(p)=\sum_{i\in\mathcal{A}}p_i\log p_i$ for all $s\in\mathcal{S}$ and $p\in\Delta(\mathcal{A})$, which coincides with the negative Shannon entropy of a probability distribution. Similar, a KL regularization adopts the choice $h_s(p)=\mathsf{KL}\left((\,\|\,p)\,\|\,p_{\mathsf{ref}}\right)$, which penalizes the distribution p that deviates from the reference p_{ref}. As another

example, a weighted ℓ_1 regularization adopts the choice $h_s(p) = \sum_{i\in\mathcal{A}} w_{s,i}p_i$ for all $s \in \mathcal{S}$ and $p \in \Delta(\mathcal{A})$, where $w_{s,i} \geq 0$ is the cost of taking action i at state s, and the regularizer $h_s(\pi(\cdot|s))$ captures the expected cost of the policy π in state s. Throughout this thesis, we impose the following assumption.

Assumption 1. *Consider an arbitrarily small constant $\zeta > 0$. For for any $s \in \mathcal{S}$, suppose that $h_s(\cdot)$ is convex and*

$$h_s(p) = \infty \qquad \text{for any } p \notin \Delta_\zeta(\mathcal{A}). \tag{3.2}$$

Following the convention in prior literature (e.g., Mei et al. [2020b]), we also define the corresponding *regularized Q-function* as follows:

$$\forall (s,a) \in \mathcal{S}\times\mathcal{A}: \qquad Q_\tau^\pi(s,a) := r(s,a) + \gamma \mathop{\mathbb{E}}_{s'\sim P(\cdot|s,a)} \big[V_\tau^\pi(s')\big]. \tag{3.3a}$$

As can be straightforwardly verified, one can also express V_τ^π in terms of Q_τ^π as

$$\forall s \in \mathcal{S}: \qquad V_\tau^\pi(s) := \mathop{\mathbb{E}}_{a\sim\pi(\cdot|s)} \Big[Q_\tau^\pi(s,a) - \tau h_s\big(\pi(\cdot\,|\,s)\big)\Big]. \tag{3.3b}$$

The optimal regularized value function $V_\tau^\star$ and the corresponding optimal policy $\pi_\tau^\star$ are defined respectively as follows:

$$\forall s \in \mathcal{S}: \qquad V_\tau^\star(s) := V_\tau^{\pi_\tau^\star}(s) = \max_\pi V_\tau^\pi(s), \qquad \pi_\tau^\star := \arg\max_\pi V_\tau^\pi. \tag{3.4}$$

It is worth noting that Puterman [2014] asserts the *existence* of an optimal policy $\pi_\tau^\star$ that achieves (3.4) simultaneously for all $s \in \mathcal{S}$. Correspondingly, we shall also define the resulting optimal regularized Q-function as

$$\forall (s,a) \in \mathcal{S}\times\mathcal{A}: \qquad Q_\tau^\star(s,a) = Q_\tau^{\pi_\tau^\star}(s,a). \tag{3.5}$$

3.1.2 Algorithm: generalized policy mirror descent

Motivated by PMD [Lan, 2023], we put forward a generalization of PMD that selects the Bregman divergence in cognizant of the regularizer in use. A thorough comparison with Lan [2023] will be provided after introducing our generalized PMD algorithm.

Review: mirror descent (MD) for the composite model. To better elucidate our algorithmic idea, let us first briefly review the design of classical mirror descent—originally proposed by Nemirovsky and Yudin [1983]—in the optimization literature. Consider the following composite model:

$$\text{minimize}_x \quad F(x) := f(x) + h(x),$$

where the objective function consists of two components. The first component is assumed to be differentiable, while the second component $h(\cdot)$ can be more general and is commonly employed to model some sort of regularizers. To solve this composite problem, one variant of mirror descent adopts the following update rule (see also Beck [2017], Duchi et al. [2010]):

$$x^{(k+1)} = \arg\min_x \left\{ f\big(x^{(k)}\big) + \big\langle \nabla f(x^{(k)}), x\big\rangle + h(x) + \frac{1}{\eta} D_h\big(x, x^{(k)}\big) \right\}, \tag{3.6}$$

where $\eta > 0$ is the learning rate or step size, and $D_h(\cdot,\cdot)$ is the Bregman divergence defined in (1.2). Note that the first term within the curly brackets of (3.6) can be safely discarded as it is a constant given $x^{(k)}$. In words, the above update rule approximates $f(x)$ via its first-order Taylor expansion $f\big(x^{(k)}\big) + \big\langle \nabla f(x^{(k)}), x\big\rangle$ at the point $x^{(k)}$, employs the Bregman divergence D_h to monitor the difference between the new iterate and the current iterate $x^{(k)}$, and attempts to optimize such (properly monitored) approximation instead. While one can further generalize the Bregman divergence to D_ω for a different generator ω, we shall restrict attention to the case with $h = \omega$ in the current paper.

The proposed algorithm. We are now ready to present the algorithm we come up with, which is an extension of the PMD algorithm [Lan, 2023]. For notational simplicity, we shall write

$$V_\tau^{(k)} \coloneqq V_\tau^{\pi^{(k)}}, \qquad Q_\tau^{(k)}(s,a) \coloneqq Q_\tau^{\pi^{(k)}}(s,a) \qquad \text{and} \qquad d_{s_0}^{(k)}(s) \coloneqq d_{s_0}^{\pi^{(k)}}(s) \tag{3.7}$$

throughout the paper, where $\pi^{(k)}$ denotes our policy estimate in the k-th iteration.

To begin with, suppose for simplicity that $h_s(\cdot)$ is differentiable everywhere. In the k-th iteration, a natural MD scheme that comes into mind for solving (3.1)—namely, maximize$_\pi V_\tau^\pi(s_0)$ for a given initial state s_0—is the following update rule:

$$\begin{aligned}
\pi^{(k+1)}(\cdot\,|\,s) &= \arg\min_{p\in\Delta(\mathcal{A})} \left\{ -\Big\langle \nabla_{\pi(\cdot|s)} V_\tau^\pi(s_0)\Big|_{\pi=\pi^{(k)}}, p\Big\rangle + \frac{\tau}{1-\gamma} d_{s_0}^{(k)}(s) h_s(p) + \frac{1}{\eta'} D_{h_s}\big(p, \pi^{(k)}(\cdot\,|\,s)\big)\right\} \\
&= \arg\min_{p\in\Delta(\mathcal{A})} \left\{ \frac{1}{1-\gamma} d_{s_0}^{(k)}(s) \Big\{ -\big\langle Q_\tau^{(k)}(s,\cdot), p\big\rangle + \tau h_s(p)\Big\} + \frac{1}{\eta'} D_{h_s}\big(p, \pi^{(k)}(\cdot\,|\,s)\big)\right\} \\
&= \arg\min_{p\in\Delta(\mathcal{A})} \left\{ -\big\langle Q_\tau^{(k)}(s,\cdot), p\big\rangle + \tau h_s(p) + \frac{1}{\eta} D_{h_s}\big(p, \pi^{(k)}(\cdot\,|\,s)\big)\right\}
\end{aligned} \tag{3.8}$$

for every state $s \in \mathcal{S}$, which is a direct application of (3.6) to our setting. Here, we start with a learning rate η', and obtain simplification by replacing η' with $\eta(1-\gamma)/d_{s_0}^{(k)}(s)$. Notably, the update strategy (3.8) is invariant to the initial state s_0, akin to natural policy gradient methods [Agarwal et al., 2020b].

This update rule is well-defined for, say, the case when h_s is the negative entropy, since the algorithm guarantees $\pi^{(k)} > 0$ all the time and hence h_s is always differentiable w.r.t. the k-th iterate (see Cen et al. [2022b]). In general, however, it is possible to encounter situations when the gradient of h_s does not exist on the boundary (e.g., when h_s represents a certain indicator function). To cope with such cases, we resort to a generalized version of Bregman divergence (e.g., Kiwiel [1997], Lan et al. [2011], Lan and Zhou [2018]). To be specific, we attempt to replace the usual Bregman divergence $D_{h_s}(p,q)$ by the following metric

$$D_{h_s}(p,q;g_s) \coloneqq h_s(p) - h_s(q) - \langle g_s, p-q\rangle \geq 0, \tag{3.9}$$

where g_s can be any vector falling within the subdifferential $\partial h_s(q)$. Here, the non-negativity condition in (3.9) follows directly from the definition of the subgradient for any convex function. The constraint on g_s can be further relaxed by exploiting the requirement $p, q \in \Delta(\mathcal{A})$. In fact, for any vector $\xi_s = g_s - c_s 1$ (with $c_s \in \mathbb{R}$ some constant and 1 the all-one vector), one can readily see that

$$\begin{aligned}
D_{h_s}(p,q;g_s) &= h_s(p) - h_s(q) - \langle g_s, p-q\rangle = h_s(p) - h_s(q) - \langle \xi_s, p-q\rangle + c_s\langle 1, p-q\rangle \\
&= h_s(p) - h_s(q) - \langle \xi_s, p-q\rangle = D_{h_s}(p,q;\xi_s),
\end{aligned} \tag{3.10}$$

where the last line is valid since $1^\top p = 1^\top q = 1$. As a result, everything boils down to identifying a vector ξ_s that falls within $\partial h_s(q)$ upon global shift.

Towards this, we propose the following iterative rule for designing such a sequence of vectors as surrogates for the subgradient of h_s:

$$\xi^{(0)}(s,\cdot) \in \partial h_s\big(\pi^{(0)}(\cdot \,|\, s)\big); \tag{3.11a}$$

$$\xi^{(k+1)}(s,\cdot) = \frac{1}{1+\eta\tau}\xi^{(k)}(s,\cdot) + \frac{\eta}{1+\eta\tau}Q_\tau^{(k)}(s,\cdot), \qquad k \geq 0, \tag{3.11b}$$

where $\xi^{(k+1)}(s,\cdot)$ is updated as a convex combination of the previous $\xi^{(k)}(s,\cdot)$ and $Q_\tau^{(k)}(s,\cdot)$, where more emphasis is put on $Q_\tau^{(k)}(s,\cdot)$ when the learning rate η is large. As asserted by the following lemma, the above vectors $\xi^{(k)}(s,\cdot)$ we construct satisfy the desired property, i.e., lying within the subdifferential of h_s under suitable global shifts. It is worth mentioning that these global shifts $\{c_s^{(k)}\}$ only serve as an aid to better understand the construction, but are not required during the algorithm updates.

Lemma 1. *For all $k \geq 0$ and every $s \in \mathcal{S}$, there exists a quantity $c_s^{(k)} \in \mathbb{R}$ such that*

$$\xi^{(k)}(s,\cdot) - c_s^{(k)}1 \in \partial h_s\big(\pi^{(k)}(\cdot \,|\, s)\big). \tag{3.12}$$

In addition, for every $s \in \mathcal{S}$, there exists a quantity $c_s^\star \in \mathbb{R}$ such that

$$\tau^{-1}Q_\tau^\star(s,\cdot) - c_s^\star 1 \in \partial h_s\big(\pi_\tau^\star(\cdot \,|\, s)\big). \tag{3.13}$$

Thus far, we have presented all crucial ingredients of our algorithm. The whole procedure is summarized in Algorithm 2, and will be referred to as *Generalized Policy Mirror Descent (GPMD)* throughout the paper. Interestingly, several well-known algorithms can be recovered as special cases of GPMD:

- When the Bregman divergence $D_{h_s}(\cdot,\cdot)$ is taken as the KL divergence, GPMD reduces to the well-renowned NPG algorithm [Kakade, 2001] when $\tau = 0$ (no regularization), and to the NPG algorithm with entropy regularization analyzed in Cen et al. [2022b] when $h_s(\cdot)$ is taken as the negative Shannon entropy.
- When $\eta = \infty$ (no divergence), GPMD reduces to regularized policy iteration in Geist et al. [2019]; in particular, GPMD reduces to the standard policy iteration algorithm if in addition τ is also 0.

Comparison with PMD [Lan, 2023]. Before continuing, let us take a moment to point out the key differences between our algorithm GPMD and the PMD algorithm proposed in Lan [2023] in terms of algorithm designs. Although the primary exposition of PMD in Lan [2023] fixes the Bregman divergence as the KL divergence, the algorithm also works in the presence of a generic Bregman divergence, whose relationship with the regularizer h_s is, however, unspecified. Furthermore, GPMD adaptively sets this term to be the Bregman divergence generated by the regularizer h_s in use, together with a carefully designed recursive update rule (cf. (3.11)) to compute surrogates for the subgradient of h_s to facilitate implementation. Encouragingly, this specific choice leads to a tailored performance analysis of GPMD, which was not present in and instead complementary with that of PMD [Lan, 2023]. In truth, our theory offers linear convergence guarantees for more general scenarios by adapting to the geometry of the regularizer h_s; details to follow momentarily.

Algorithm 2: PMD with generalized Bregman divergence (GPMD)

Input: initial policy iterate $\pi^{(0)}$, learning rate $\eta > 0$.

Initialize $\xi^{(0)}$ so that $\xi^{(0)}(s,\cdot) \in \partial h_s\big(\pi^{(0)}(\cdot|s)\big)$ for all $s \in \mathcal{S}$.

for $k = 0, 1, \cdots$, **do**

For every $s \in \mathcal{S}$, set

$$\pi^{(k+1)}(\cdot|s) = \arg\min_{p \in \Delta(\mathcal{A})} \left\{ -\big\langle Q_\tau^{(k)}(s,\cdot), p\big\rangle + \tau h_s(p) + \frac{1}{\eta} D_{h_s}\big(p, \pi^{(k)}(\cdot|s); \xi^{(k)}\big) \right\}, \tag{3.14a}$$

where

$$D_{h_s}\big(p, q; \xi\big) \coloneqq h_s(p) - h_s(q) - \big\langle \xi(s,\cdot), p - q\big\rangle. \tag{3.14b}$$

For every $(s,a) \in \mathcal{S} \times \mathcal{A}$, compute

$$\xi^{(k+1)}(s,a) = \frac{1}{1+\eta\tau}\xi^{(k)}(s,a) + \frac{\eta}{1+\eta\tau} Q_\tau^{(k)}(s,a). \tag{3.14c}$$

3.2 Main results

This section presents our convergence guarantees for the GPMD method presented in Algorithm 2. We shall start with the idealized case assuming that the update rule can be precisely implemented, and then discuss how to generalize it to the scenario with imperfect policy evaluation.

3.2.1 Convergence of exact GPMD

To start with, let us pin down the convergence behavior of GPMD, assuming that accurate evaluation of the policy $Q_\tau^{(k)}$ is available and the subproblem (3.14a) can be solved perfectly. Here and below, we shall refer to the algorithm in this case as exact GPMD. Encouragingly, exact GPMD provably achieves global linear convergence from an arbitrary initialization, as asserted by the following theorem.

Theorem 3 (Exact GPMD)**.** *Suppose that Assumption 1 holds. Consider any learning rate $\eta > 0$, and set $\alpha := \frac{1}{1+\eta\tau}$. Then the iterates of Algorithm 2 satisfy*

$$\big\|Q_\tau^\star - Q_\tau^{(k+1)}\big\|_\infty \le \gamma\big(1 - (1-\alpha)(1-\gamma)\big)^k C_1, \tag{3.15a}$$

$$\big\|V_\tau^\star - V_\tau^{(k+1)}\big\|_\infty \le (\gamma+2)\big(1 - (1-\alpha)(1-\gamma)\big)^k C_1, \tag{3.15b}$$

for all $k \ge 0$, where $C_1 \coloneqq \|Q_\tau^\star - Q_\tau^{(0)}\|_\infty + 2\alpha\|Q_\tau^\star - \tau\xi^{(0)}\|_\infty$.

In addition, if h_s is 1-strongly convex w.r.t. the ℓ_1 norm for some $s \in \mathcal{S}$, then one further has

$$\big\|\pi_\tau^\star(s) - \pi_\tau^{(k+1)}(s)\big\|_1 \le \tau^{-1}\big(1 - (1-\alpha)(1-\gamma)\big)^k C_1, \qquad k \ge 0. \tag{3.16}$$

Our theorem confirms the fast global convergence of the GPMD algorithm, in terms of both the resulting regularized Q-value (if $h_s(\cdot)$ is convex) and the policy estimate (if $h_s(\cdot)$ is strongly convex). In summary, it takes GPMD no more than

$$\frac{1}{(1-\alpha)(1-\gamma)} \log\frac{C_1}{\varepsilon} = \frac{1+\eta\tau}{\eta\tau(1-\gamma)} \log\frac{C_1}{\varepsilon} \tag{3.17a}$$

iterations to converge to an ε-optimal regularized Q-function (in the ℓ_∞ sense), or

$$\frac{1}{(1-\alpha)(1-\gamma)}\log\frac{C_1}{\varepsilon\tau} = \frac{1+\eta\tau}{\eta\tau(1-\gamma)}\log\frac{C_1}{\varepsilon\tau} \tag{3.17b}$$

iterations to yield an ε-approximation (w.r.t. the ℓ_1 norm error) of $\pi_\tau^\star$. The iteration complexity (3.17) is nearly dimension-free—namely, depending at most logarithmically on the dimension of the state-action space —making it scalable to large-dimensional problems.

Comparison with Lan [2023, Theorems 1-3]. To make clear our contributions, it is helpful to compare Theorem 3 with the theory for the state-of-the-art algorithm PMD in Lan [2023].

- *Linear convergence for convex regularizers under constant learning rates.* Suppose that constant learning rates are adopted for both GPMD and PMD. Our finding reveals that GPMD enjoys global linear convergence—in terms of both $\|Q_\tau^\star - Q_\tau^{(k+1)}\|_\infty$ and $\|V_\tau^\star - V_\tau^{(k+1)}\|_\infty$—even when the regularizer $h_s(\cdot)$ is only convex but not strongly convex. In contrast, Lan [2023, Theorem 2] provided only sublinear convergence guarantees (with an iteration complexity proportional to $1/\varepsilon$) for the case with convex regularizers, provided that constant learning rates are adopted.[1]

- *A full range of learning rates.* Theorem 3 reveals linear convergence of GPMD for a full range of learning rates, namely, our result is applicable to any $\eta > 0$. In comparison, linear convergence was established in Lan [2023] only when the learning rates are sufficiently large and when h_s is 1-strongly convex w.r.t. the KL divergence. Consequently, the linear convergence results in Lan [2023] do not extend to several widely used regularizers such as negative Tsallis entropy and log-barrier functions (even after scaling), which are, in contrast, covered by our theory. It is worth noting that the case with small-to-medium learning rates is often more challenging to cope with in theory, given that its dynamics could differ drastically from that of regularized policy iteration.

- *Further comparison of rates under large learning rates.* [Lan, 2023, Theorem 1] achieves a contraction rate of γ when the regularizer is strongly convex and the step size satisfies $\eta \geq \frac{1-\gamma}{\gamma\tau}$, while the contraction rate of GPMD is $1 - \frac{\eta\tau}{1+\eta\tau}(1-\gamma)$ under the full range of the step size, which is slower but approaches the contraction rate γ of PMD as η goes to infinity. Therefore, in the limit $\eta \to \infty$, both GPMD and PMD achieve the contraction rate γ. As soon as $\eta \geq 1/\tau$, their iteration complexities are on the same order.

3.2.2 Convergence of approximate GPMD

In reality, however, it is often the case that GPMD cannot be implemented in an exact manner, either because perfect policy evaluation is unavailable or because the subproblem (3.14a) cannot be solved exactly. To accommodate these practical considerations, this subsection generalizes our previous result by permitting inexact policy evaluation and non-zero optimization error in solving (3.14a). The following assumptions make precise this imperfect scenario.

Assumption 2 (Policy evaluation error)**.** *Suppose for any $k \geq 0$, we have access to an estimate $\hat{Q}_\tau^{(k)}$ obeying*

$$\left\|\hat{Q}_\tau^{(k)} - Q_\tau^{(k)}\right\|_\infty \leq \varepsilon_{\mathsf{eval}}. \tag{3.18}$$

[1] In fact, Lan [2023, Theorem 3] suggests using a vanishing strongly convex regularization, as well as a corresponding increasing sequence of learning rates, in order to enable linear convergence for non-strongly-convex regularizers.

Assumption 3 (Subproblem optimization error). *Consider any policy π and any vector $\xi \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$. Define*

$$f_s(p;\pi,\xi) \coloneqq -\big\langle Q(s,\cdot),p\big\rangle + \tau h_s(p) + \frac{1}{\eta} D_{h_s}\big(p,\pi(\cdot\,|\,s);\xi(s,\cdot)\big),$$

where $D_{h_s}(p,q;\xi)$ is defined in (3.9). Suppose there exists an oracle $G_{s,\varepsilon_{\mathsf{opt}}}(Q,\pi,\xi)$, which is capable of returning $\pi'(\cdot\,|\,s)$ such that

$$f_s\big(\pi'(\cdot\,|\,s);\pi,\xi\big) \le \min_{p\in\Delta(\mathcal{A})} f_s(p;\pi,\xi) + \varepsilon_{\mathsf{opt}}. \tag{3.19}$$

Note that the oracle in Assumption 3 can be implemented efficiently in practice via various first-order methods [Beck, 2017]. Under Assumptions 2 and 3, we can modify Algorithm 2 by replacing $\{Q_\tau^{(k)}\}$ with the estimate $\{\hat{Q}_\tau^{(k)}\}$, and invoking the oracle $G_{s,\varepsilon_{\mathsf{opt}}}(Q,\pi,\xi)$ to solve the subproblem (3.14a) approximately. The whole procedure, which we shall refer to as approximate GPMD, is summarized in Algorithm 3.

Algorithm 3: Approximate PMD with generalized Bregman divergence (Approximate GPMD)

Input: initial policy $\pi^{(0)}$, learning rate $\eta > 0$.
Initialize $\hat{\xi}^{(0)}(s) \in \partial h_s\big(\pi^{(0)}(\cdot\,|\,s)\big)$ for all $s \in \mathcal{S}$.
for $k = 0, 1, \cdots$, **do**
  For every $s \in \mathcal{S}$, invoke the oracle to obtain (cf. (3.19))

$$\pi^{(k+1)}(s) = G_{s,\varepsilon_{\mathsf{opt}}}\big(\hat{Q}_\tau^{(k)}, \pi^{(k)}, \hat{\xi}^{(k)}\big). \tag{3.20}$$

  For every $(s,a) \in \mathcal{S}\times\mathcal{A}$, compute

$$\hat{\xi}^{(k+1)}(s,a) = \frac{1}{1+\eta\tau}\hat{\xi}^{(k)}(s,a) + \frac{\eta}{1+\eta\tau}\hat{Q}_\tau^{(k)}(s,a). \tag{3.21}$$

The following theorem uncovers that approximate GPMD converges linearly—at the same rate as exact GPMD—before an error floor is hit.

Theorem 4 (Approximate GPMD). *Suppose that Assumptions 1, 2 and 3 hold. Consider any learning rate $\eta > 0$. Then the iterates of Algorithm 3 satisfy*

$$\|Q_\tau^\star - Q_\tau^{(k+1)}\|_\infty \le \gamma\left[\big(1-(1-\alpha)(1-\gamma)\big)^k C_1 + C_2\right], \tag{3.22a}$$

$$\|V_\tau^\star - V_\tau^{(k+1)}\|_\infty \le (\gamma+2)\left[\big(1-(1-\alpha)(1-\gamma)\big)^k C_1 + C_2\right] + (1-\alpha)\varepsilon_{\mathsf{opt}}, \tag{3.22b}$$

where $\alpha \coloneqq \frac{1}{1+\eta\tau}$, C_1 is defined in Theorem 3, and

$$C_2 \coloneqq \frac{1}{1-\gamma}\left[\left(2+\frac{2\gamma}{(1-\gamma)(1-\alpha)}\right)\varepsilon_{\mathsf{eval}} + \left(1+\frac{2\gamma}{(1-\gamma)(1-\alpha)}\right)\varepsilon_{\mathsf{opt}}\right].$$

In addition, if h_s is 1-strongly convex w.r.t. the ℓ_1 norm for any $s \in \mathcal{S}$, then we can further

obtain

$$\|Q_\tau^\star - Q_\tau^{(k+1)}\|_\infty \le \gamma\left[\big(1-(1-\alpha)(1-\gamma)\big)^k C_1 + C_3\right], \tag{3.23a}$$

$$\|V_\tau^\star - V_\tau^{(k+1)}\|_\infty \le (\gamma+2)\left[\big(1-(1-\alpha)(1-\gamma)\big)^k C_1 + C_3\right] + (1-\alpha)\varepsilon_{\mathsf{opt}}, \tag{3.23b}$$

$$\left\|\pi_\tau^\star(\cdot\,|\,s) - \pi^{(k+1)}(\cdot\,|\,s)\right\|_1 \le \tau^{-1}\left[\big(1-(1-\alpha)(1-\gamma)\big)^k C_1 + C_3\right] + \sqrt{\frac{2\eta\varepsilon_{\mathsf{opt}}}{1+\eta\tau}}, \tag{3.23c}$$

where

$$C_3 \coloneqq \frac{1}{1-\gamma}\left[\left(2+\frac{\varepsilon_{\mathsf{eval}}\gamma}{\tau(1-\gamma)}\right)\varepsilon_{\mathsf{eval}} + \left(1+\frac{4\gamma}{(1-\gamma)(1-\alpha)}\right)\varepsilon_{\mathsf{opt}}\right]. \tag{3.24}$$

In the special case where $\varepsilon_{\mathsf{opt}} = 0$ and $\eta = \infty$, Algorithm 3 reduces to regularized policy iteration, and the convergence result can be simplified as follows

$$\left\|Q_\tau^\star - Q_\tau^{(k)}\right\|_\infty \le \gamma^k \left\|Q_\tau^\star - Q_\tau^{(0)}\right\|_\infty + \frac{2\gamma\varepsilon_{\mathsf{eval}}}{(1-\gamma)`2}.$$

In particular, when h_s is taken as the negative entropy, our result strengthens the prior result established in Cen et al. [2022b] for approximate entropy-regularized NPG method with $\varepsilon_{\mathsf{opt}} = 0$ over a wide range of learning rates. Specifically, the error bound in Cen et al. [2022b] reads $\gamma \cdot \frac{\varepsilon_{\mathsf{eval}}}{1-\gamma}\left(2+\frac{2\gamma}{\eta\tau}\right)$, where the second term in the bracket scales inversely with respect to η and therefore grows unboundedly as η approaches 0. In contrast, (3.23) and (3.24) suggest a bound $\gamma \cdot \frac{\varepsilon_{\mathsf{eval}}}{1-\gamma}\left(2+\frac{\varepsilon_{\mathsf{eval}}\gamma}{\tau(1-\gamma)}\right)$, which is independent of the learning rate η in use and thus prevents the error bound from blowing up when the learning rate approaches 0. Indeed, our result improves over the prior art Cen et al. [2022b] whenever $\eta \le \frac{2(1-\gamma)}{\varepsilon_{\mathsf{eval}}}$.

Remark 2 (Sample complexities)**.** *One might naturally ask how many samples are sufficient to learn an ε-optimal regularized Q-function, by leveraging sample-based policy evaluation algorithms in GPMD. Notice that it is straightforward to consider an expected version of Assumption 2 as following:*

$$\begin{cases} \mathbb{E}\big[\big\|\hat{Q}_\tau^{(k)} - Q_\tau^{(k)}\big\|_\infty\big] & \le \varepsilon_{\mathsf{eval}}; \\ \mathbb{E}\big[\big\|\hat{Q}_\tau^{(k)} - Q_\tau^{(k)}\big\|_\infty^2\big] & \le \varepsilon_{\mathsf{eval}}^2, \end{cases}$$

where the expectation is with respect to the randomness in policy evaluation, then the convergence results in Theorem 4 apply to $\mathbb{E}\big[\|Q_\tau^\star - Q_\tau^{(k+1)}\|_\infty\big]$ and $\mathbb{E}\big[\big\|\pi_\tau^\star(\cdot\,|\,s) - \pi_\tau^{(k+1)}(\cdot\,|\,s)\big\|_1\big]$ instead. This randomized version makes it immediately amenable to combine with, e.g., the rollout-based policy evaluators in Lan [2023, Section 5.1] to obtain (possibly crude) bounds on the sample complexity. We omit these straightforward developments.

Roughly speaking, approximate GPMD is guaranteed to converge linearly to an error bound that scales linearly in both the policy evaluation error $\varepsilon_{\mathsf{eval}}$ and the optimization error $\varepsilon_{\mathsf{opt}}$, thus confirming the stability of our algorithm vis-à-vis imperfect implementation of the algorithm. As before, our theory improves upon prior works by demonstrating linear convergence for a full range of learning rates even in the absence of strong convexity and smoothness.

3.3 Discussion

The present paper has introduced a generalized framework of policy optimization tailored to regularized RL problems. We have proposed a generalized policy mirror descent (GPMD) algorithm

that achieves dimension-free linear convergence, which covers an entire range of learning rates and accommodates convex and possibly nonsmooth regularizers. Numerical experiments have been conducted to demonstrate the utility of the proposed GPMD algorithm. Our approach opens up a couple of future directions that are worthy of further exploration. For example, the current work restricts its attention to convex regularizers and tabular MDPs; it is of paramount interest to develop policy optimization algorithms when the regularizers are nonconvex and when sophisticated policy parameterization—including function approximation—is adopted. Understanding the sample complexities of the proposed algorithm—when the policies are evaluated using samples collected over an online trajectory—is crucial in sample-constrained scenarios and is left for future investigation. Furthermore, it might be worthwhile to extend the proposed algorithm to accommodate multi-agent RL, with a representative example being regularized multi-agent Markov games [Cen et al., 2021, Zhao et al., 2022, Cen et al., 2022a, 2023].

Part II

Policy optimization for multi-agent Systems

Chapter 4

Two-player Zero-sum Matrix Games

In this chapter, we consider a two-player zero-sum game with bilinear objective and probability simplex constraints, and demonstrate the positive role of entropy regularization in solving this problem. Throughout this thesis, let $\mathcal{A} = \{1, \ldots, m\}$ and $\mathcal{B} = \{1, \ldots, n\}$ be the action spaces of each player. For more details and entire analysis, please refer to Cen et al. [2021].

4.1 Background and problem formulation

Zero-sum two-player matrix game. The focal point of this section is a constrained two-player zero-sum matrix game, which can be formulated as the following min-max problem (or saddle point optimization problem):

$$\max_{\mu \in \Delta(\mathcal{A})} \min_{\nu \in \Delta(\mathcal{B})} f(\mu, \nu) \coloneqq \mu^\top A \nu, \tag{4.1}$$

where $A \in \mathbb{R}^{m \times n}$ denotes the payoff matrix, $\mu \in \Delta(\mathcal{A})$ and $\nu \in \Delta(\mathcal{B})$ stand for the mixed/randomized policies of each player, defined respectively as distributions over the probability simplex $\Delta(\mathcal{A})$ and $\Delta(\mathcal{B})$. It is well known since Neumann [1928] that the max and min operators in (4.1) can be exchanged without affecting the solution. A pair of policies $(\mu^\star, \nu^\star)$ is said to be a *Nash equilibrium (NE)* of (4.1) if

$$f(\mu^\star, \nu) \geq f(\mu^\star, \nu^\star) \geq f(\mu, \nu^\star) \qquad \text{for all } (\mu, \nu) \in \Delta(\mathcal{A}) \times \Delta(\mathcal{B}). \tag{4.2}$$

In words, the NE corresponds to when both players play their best-response strategies against their respective opponents.

Entropy-regularized zero-sum two-player matrix game. There is no shortage of scenarios where the payoff matrix A might not be known perfectly. In an attempt to accommodate imperfect knowledge of A, McKelvey and Palfrey [1995] proposed a seminal extension to the Nash equilibrium called the *quantal response equilibrium (QRE)* when the payoffs are perturbed by Gumbel-distributed noise. Formally, this amounts to solving the following matrix game with entropy regularization [Mertikopoulos and Sandholm, 2016]:

$$\max_{\mu \in \Delta(\mathcal{A})} \min_{\nu \in \Delta(\mathcal{B})} f_\tau(\mu, \nu) \coloneqq \mu^\top A \nu + \tau \mathcal{H}(\mu) - \tau \mathcal{H}(\nu), \tag{4.3}$$

where $\mathcal{H}(\pi) \coloneqq -\sum_i \pi_i \log(\pi_i)$ denotes the Shannon entropy of a distribution π, and $\tau \geq 0$ is the regularization parameter. As is well known, the optimal solution $(\mu_\tau^\star, \nu_\tau^\star)$ to (4.3), dubbed as the

QRE, is unique whenever $\tau > 0$ (due to the presence of strong concavity/convexity), which satisfies the following fixed point equations:

$$\begin{cases} \mu_\tau^\star(a) = \frac{\exp([A\nu_\tau^\star]_a/\tau)}{\sum_{a=1}^m \exp([A\nu_\tau^\star]_a/\tau)} \propto \exp([A\nu_\tau^\star]_a/\tau), & \text{for all } a \in \mathcal{A}, \\ \nu_\tau^\star(b) = \frac{\exp(-[A^\top\mu_\tau^\star]_b/\tau)}{\sum_{b=1}^n \exp(-[A^\top\mu_\tau^\star]_b/\tau)} \propto \exp(-[A^\top\mu_\tau^\star]_b/\tau), & \text{for all } b \in \mathcal{B}. \end{cases} \tag{4.4}$$

Goal. We aim to efficiently compute the QRE of the entropy-regularized matrix game in a decentralized manner, and investigate how an efficient solver of QRE can be leveraged to find a NE of the unregularized matrix game (4.1). Namely, we only assume access to "first-order information" as opposed to full knowledge of the payoff matrix A or the actions of the opponent. The information received by each player is formally described in the following sampling oracle.

Definition 1 (Sampling oracle for matrix games)**.** *For any policy pair (μ, ν) and payoff matrix A, the sampling oracle returns the exact values of $\mu^\top A$ and $A\nu$.*

Additional notation. For notational convenience, we let ζ represent the concatenation of $\mu \in \mathbb{R}^{|\mathcal{A}|}$ and $\nu \in \mathbb{R}^{|\mathcal{B}|}$, namely, $\zeta = (\mu, \nu)$. The solution to (4.3), which is specified in (4.4), is denoted by $\zeta_\tau^\star = (\mu_\tau^\star, \nu_\tau^\star)$. For any $\zeta = (\mu, \nu)$ and $\zeta' = (\mu', \nu')$, we shall often abuse the notation and let

$$\mathsf{KL}\left(\zeta \,\|\, \zeta'\right) = \mathsf{KL}\left(\mu \,\|\, \mu'\right) + \mathsf{KL}\left(\nu \,\|\, \nu'\right).$$

The duality gap of the entropy-regularized matrix game (4.3) at $\zeta = (\mu, \nu)$ is defined as

$$\mathsf{DualGap}_\tau(\zeta) = \max_{\mu' \in \Delta(\mathcal{A})} f_\tau(\mu', \nu) - \min_{\nu' \in \Delta(\mathcal{B})} f_\tau(\mu, \nu') \tag{4.5}$$

which is clearly nonnegative and $\mathsf{DualGap}_\tau(\zeta_\tau^\star) = 0$. Similarly, let the optimality gap of the entropy-regularized matrix game (4.3) at $\zeta = (\mu, \nu)$ be $\mathsf{OptGap}(\zeta) = \left|f_\tau(\mu, \nu) - f_\tau(\mu_\tau^\star, \nu_\tau^\star)\right|$.

4.2 Proposed extragradient methods: PU and OMWU

To begin, assume we are given a pair of policies $z_1 \in \mathcal{A}$, $z_2 \in \mathcal{B}$ employed by each player respectively. If we proceed with fictitious play, i.e. player 1 (resp. player 2) aims to optimize its own policy by assuming the opponent's policy is fixed as z_2 (resp. z_1), the saddle-point optimization problem (4.3) is then decoupled into two independent min/max optimization problems:

$$\max_{\mu \in \Delta(\mathcal{A})} \ \mu^\top A z_2 + \tau\mathcal{H}(\mu) - \tau\mathcal{H}(z_2) \qquad \text{and} \qquad \min_{\nu \in \Delta(\mathcal{B})} \ z_1^\top A \nu + \tau\mathcal{H}(z_1) - \tau\mathcal{H}(\nu),$$

which are naturally solved via mirror descent / ascent with KL divergence. Specifically, one step of mirror descent / ascent takes the form

$$\begin{cases} \mu^{(t+1)} = \arg\max_{\mu \in \Delta(\mathcal{A})} \ (Az_2 - \tau \log \mu^{(t)})^\top \mu - \frac{1}{\eta}\mathsf{KL}\left(\mu \,\|\, \mu^{(t)}\right) \\ \nu^{(t+1)} = \arg\min_{\nu \in \Delta(\mathcal{B})} \ (A^\top z_1 + \tau \log \nu^{(t)})^\top \nu + \frac{1}{\eta}\mathsf{KL}\left(\nu \,\|\, \nu^{(t)}\right) \end{cases},$$

where η is the learning rate, or equivalently

$$\begin{cases} \mu^{(t+1)}(a) \propto \mu^{(t)}(a)^{1-\eta\tau} \exp(\eta[Az_2]_a), & \text{for all } a \in \mathcal{A}, \\ \nu^{(t+1)}(b) \propto \nu^{(t)}(b)^{1-\eta\tau} \exp(-\eta[A^\top z_1]_b), & \text{for all } b \in \mathcal{B}. \end{cases} \tag{4.6}$$

The above update rule forms the basis of our algorithm design.

Motivation: a form of implicit updates with linear convergence. To begin with, we select the policy pair $(z_1, z_2) = \zeta^{(t+1)} := (\mu^{(t+1)}, \nu^{(t+1)})$ as the solution to the following equations, and call the conceptual update rule as the Implicit Update (IU) method:

$$\text{Implicit Update:} \qquad \begin{cases} \mu^{(t+1)}(a) \propto \mu^{(t)}(a)^{1-\eta\tau} \exp(\eta[A\nu^{(t+1)}]_a), & \text{for all } a \in \mathcal{A}, \\ \nu^{(t+1)}(b) \propto \nu^{(t)}(b)^{1-\eta\tau} \exp(-\eta[A^\top \mu^{(t+1)}]_b), & \text{for all } b \in \mathcal{B}. \end{cases} \tag{4.7}$$

Though unrealistic — since it uses the future updates and denies closed-form solutions — it leads to a one-step convergence to the QRE when $\eta = 1/\tau$ (see the optimality condition in (4.4)). Encouragingly, we have the following linear convergence guarantee of IU when adopting a general learning rate.

Proposition 1 (Linear convergence of IU)**.** *Assume $0 < \eta \le 1/\tau$, then for all $t \ge 0$, the iterates $\zeta^{(t)} := (\mu^{(t)}, \nu^{(t)})$ of the IU method in (4.7) satisfy*

$$\mathsf{KL}\left(\zeta_\tau^\star \,\|\, \zeta^{(t)}\right) \le (1-\eta\tau)^t \mathsf{KL}\left(\zeta_\tau^\star \,\|\, \zeta^{(0)}\right).$$

In words, the IU method achieves an appealing linear rate of convergence that is independent of the problem dimension. Motivated by this observation, we seek to design algorithms where the policies (z_1, z_2) employed in (4.6) serve as good predictions of $(\mu^{(t+1)}, \nu^{(t+1)})$, such that the resulting algorithms are both practical and retain the appealing convergence rate of IU.

Proposed algorithms. We propose two extragradient algorithms for solving the entropy-regularized matrix game, namely the *Predictive Update (PU)* method and the *Optimistic Multiplicative Weights Update (OMWU)* method, where the latter is adapted from Rakhlin and Sridharan [2013]. Detailed procedures can be found in Algorithm 4 and Algorithm 5, respectively. On a high level, both algorithms maintain two intertwined sequences $\{(\mu^{(t)}, \nu^{(t)})\}_{t\ge 0}$ and $\{(\bar\mu^{(t)}, \bar\nu^{(t)})\}_{t\ge 0}$, and in each iteration $t = 0, 1, \ldots$, proceed in two steps:

- The midpoint $(\bar\mu^{(t+1)}, \bar\nu^{(t+1)})$ serves as a prediction of $(\mu^{(t+1)}, \nu^{(t+1)})$ by running one step of mirror descent / ascent (cf. (4.6)) from either $(z_1, z_2) = (\mu^{(t)}, \nu^{(t)})$ (for PU) or $(z_1, z_2) = (\bar\mu^{(t)}, \bar\nu^{(t)})$ (for OMWU).
- The update of $(\mu^{(t+1)}, \nu^{(t+1)})$ then mimics the implicit update (4.7) using the prediction $(\bar\mu^{(t+1)}, \bar\nu^{(t+1)})$ obtained above.

When the proposed algorithms converge, both $(\mu^{(t)}, \nu^{(t)})$ and $(\bar\mu^{(t)}, \bar\nu^{(t)})$ converge to the same point. The two players are completely symmetric and adopt the same learning rate, and require *only* first-order information provided by the sampling oracle. While the two algorithms resemble each other in many aspects, a key difference lies in the query and use of the sampling oracle: in each iteration, OMWU makes a single call to the sampling oracle for gradient evaluation, while PU calls the sampling oracle twice. It is worth noting that, when $\tau = 0$ (i.e., no entropy regularization is enforced), the OMWU method in Algorithm 5 reduces to the method analyzed in Rakhlin and Sridharan [2013], Daskalakis and Panageas [2019], Wei et al. [2021a] without entropy regularization.

Remark 3. *It is worth highlighting that the proposed algorithms are* different *from the mirror prox algorithm [Nemirovski, 2004] or the optimistic mirror descent method [Mertikopoulos et al., 2018a], as the extragradient is only applied to the bilinear term but not the entropy regularization term. This seemingly small, but important, difference leads to a more concise closed-form update rule and a cleaner analysis, as shall be seen momentarily.*

Algorithm 4: The PU method

initialization: $\mu^{(0)}$, $\nu^{(0)}$.

parameters: learning rate η_t.

for $t = 0, 1, 2, \cdots$ **do**

Update $\bar{\mu}$ and $\bar{\nu}$ according to

$$\begin{cases} \bar{\mu}^{(t+1)}(a) \propto \mu^{(t)}(a)^{1-\eta_t\tau} \exp(\eta_t [A\nu^{(t)}]_a), \\ \bar{\nu}^{(t+1)}(b) \propto \nu^{(t)}(b)^{1-\eta_t\tau} \exp(-\eta_t [A^\top \mu^{(t)}]_b). \end{cases}$$

Update μ and ν according to

$$\begin{cases} \mu^{(t+1)}(a) \propto \mu^{(t)}(a)^{1-\eta_t\tau} \exp(\eta_t [A\bar{\nu}^{(t+1)}]_a), \\ \nu^{(t+1)}(b) \propto \nu^{(t)}(b)^{1-\eta_t\tau} \exp(-\eta_t [A^\top \bar{\mu}^{(t+1)}]_b). \end{cases}$$

Algorithm 5: The OMWU method

initialization: $\mu^{(0)} = \bar{\mu}^{(0)}$, $\nu^{(0)} = \bar{\nu}^{(0)}$.

parameters: learning rate η_t.

for $t = 0, 1, 2, \cdots$ **do**

Update $\bar{\mu}$ and $\bar{\nu}$ according to

$$\begin{cases} \bar{\mu}^{(t+1)}(a) \propto \mu^{(t)}(a)^{1-\eta_t\tau} \exp(\eta_t [A\bar{\nu}^{(t)}]_a), \\ \bar{\nu}^{(t+1)}(b) \propto \nu^{(t)}(b)^{1-\eta_t\tau} \exp(-\eta_t [A^\top \bar{\mu}^{(t)}]_b). \end{cases}$$

Update μ and ν according to

$$\begin{cases} \mu^{(t+1)}(a) \propto \mu^{(t)}(a)^{1-\eta_t\tau} \exp(\eta_t [A\bar{\nu}^{(t+1)}]_a), \\ \nu^{(t+1)}(b) \propto \nu^{(t)}(b)^{1-\eta_t\tau} \exp(-\eta_t [A^\top \bar{\mu}^{(t+1)}]_b). \end{cases}$$

4.3 Last-iterate linear convergence guarantees

We are now positioned to present our main theorem concerning the last-iterate convergence of PU and OMWU for solving (4.3).

Theorem 5 (Last-iterate convergence of PU and OMWU)**.** *Suppose that the learning rates $\eta_t = \eta = \eta_{\mathsf{PU}}$ of PU in Algorithm 4 and $\eta_t = \eta = \eta_{\mathsf{OMWU}}$ of OMWU in Algorithm 5 satisfy*

$$0 < \eta_{\mathsf{PU}} \le \frac{1}{\tau + 2\|A\|_\infty}, \quad \text{and} \quad 0 < \eta_{\mathsf{OMWU}} \le \min\left\{\frac{1}{2\tau + 2\|A\|_\infty}, \frac{1}{4\|A\|_\infty}\right\}. \tag{4.8}$$

Then for any $t \ge 0$, the iterates $\zeta^{(t)} = (\mu^{(t)}, \nu^{(t)})$ and $\bar{\zeta}^{(t)} = (\bar{\mu}^{(t)}, \bar{\nu}^{(t)})$ of both PU and OMWU achieve

- **Linear convergence of policies in KL divergence and entrywise log-ratios:**

$$\max\left\{\mathsf{KL}\big(\zeta_\tau^\star \,\|\, \zeta^{(t)}\big),\ \tfrac{1}{2}\mathsf{KL}\left(\zeta_\tau^\star \,\|\, \bar{\zeta}^{(t+1)}\right)\right\} \le (1-\eta\tau)^t \mathsf{KL}\big(\zeta_\tau^\star \,\|\, \zeta^{(0)}\big), \tag{4.9a}$$

$$\left\|\log\frac{\zeta^{(t)}}{\zeta_\tau^\star}\right\|_\infty \le 2(1-\eta\tau)^t \left\|\log\frac{\zeta^{(0)}}{\zeta_\tau^\star}\right\|_\infty + \frac{8\|A\|_\infty}{\tau}(1-\eta\tau)^{t/2}\mathsf{KL}\big(\zeta_\tau^\star \,\|\, \zeta^{(0)}\big)^{1/2}. \tag{4.9b}$$

- **Linear convergence of values in optimality and duality gaps:**

$$\mathsf{OptGap}_\tau(\bar{\zeta}^{(t)}) \le 3\eta^{-1}(1-\eta\tau)^t \mathsf{KL}\big(\zeta_\tau^\star \,\|\, \zeta^{(0)}\big), \tag{4.9c}$$

$$\mathsf{DualGap}_\tau(\bar{\zeta}^{(t)}) \le \left(\eta^{-1} + 2\tau^{-1}\|A\|_\infty^2\right)(1-\eta\tau)^{t-1}\mathsf{KL}\big(\zeta_\tau^\star \,\|\, \zeta^{(0)}\big). \tag{4.9d}$$

Remark 4. *To further understand the term $\mathsf{KL}\big(\zeta_\tau^\star \,\|\, \zeta^{(0)}\big)$ in (4.9), setting $\mu^{(0)}$ and $\nu^{(0)}$ to be uniform policies leads to a universal bound*

$$\mathsf{KL}\big(\zeta_\tau^\star \,\|\, \zeta^{(0)}\big) = \log|\mathcal{A}| + \log|\mathcal{B}| - \mathcal{H}(\mu_\tau^\star) - \mathcal{H}(\nu_\tau^\star) \le \log|\mathcal{A}| + \log|\mathcal{B}|$$

regardless of $\zeta_\tau^\star = (\mu_\tau^\star, \nu_\tau^\star)$.

Remark 5. *Similar results continue to hold even when the two players use different regularization parameters $\tau_\mu, \tau_\nu > 0$ in (4.3), as long as the regularization parameter τ is replaced by $\max\{\tau_\mu, \tau_\nu\}$ in the upper bounds of the learning rate, and the contraction parameter is replaced by $1 - \min\{\tau_\mu, \tau_\nu\}\eta$.*

Theorem 5 characterizes the convergence of the *last-iterates* $\zeta^{(t)}$ and $\bar{\zeta}^{(t)}$ of PU and OMWU as long as the learning rate lies within the specified ranges. While PU doubles the number of calls to the sampling oracle, it also allows roughly as large as twice the learning rate compared with OMWU (cf. (4.8)). Compared with the vast literature analyzing the average-iterate performance of variants of extragradient methods [Daskalakis et al., 2011, Rakhlin and Sridharan, 2013], our results contribute towards characterizing the last-iterate convergence of multiplicative update methods in the presence of entropy regularization and simplex constraints, which to the best of our knowledge, are the first of its kind. Several remarks are in order.

- **Linear convergence to QRE.** To achieve an ε-accurate estimate of the QRE in terms of the KL divergence, the bound (4.9a) tells that it is sufficient to take

$$\frac{1}{\eta\tau}\log\left(\frac{\log|\mathcal{A}| + \log|\mathcal{B}|}{\varepsilon}\right)$$

  iterations using either PU or OMWU. Notably, this iteration complexity does not depend on any hidden constants and only depends double logarithmically on the cardinality of action spaces, which is almost dimension-free. Maximizing the learning rate, the iteration complexity is bounded by $(1 + \|A\|_\infty/\tau)\log(1/\varepsilon)$ (modulo log factors), which only depends on the ratio $\|A\|_\infty/\tau$.

- **Entrywise error of the policy log-ratios.** Both PU and OMWU enjoy strong entrywise guarantees in the sense we can guarantee the convergence of the ℓ_∞ norm of the log-ratios between the learned policy pair and the QRE at the same dimension-free linear rate (cf. (4.9b)), which suggests the policy pair converges in a somewhat uniform manner across the entire action space.

- **Linear convergence of optimality and duality gaps.** Our theorem also establishes the last-iterate convergence of the game values in terms of the optimality gap (cf. (4.9c)) and the duality gap (cf. (4.9d)) for both PU and OMWU. In particular, as will be seen, bounding the optimality gap of matrix games turns out to be the key enabler for generalizing our algorithms to Markov games, and bounding the duality gap allows to directly translate our results to finding a NE of unregularized matrix games.

Figure 4.1 illustrates the performance of the proposed PU and OMWU methods for solving randomly generated entropy-regularized matrix games. It is evident that both algorithms converge linearly, and achieve faster convergence rates when the regularization parameter increases.

Last-iterate convergence to approximate NE. The entropy-regularized matrix game can be thought as a smooth surrogate of the unregularized matrix game (4.1); in particular, it is possible to find an ε-NE by setting τ sufficiently small in (4.3). According to [Zhang et al., 2020c, Definition 2.1], a policy pair $\zeta = (\mu, \nu)$ is an ε-NE if it satisfies

$$\mathsf{DualGap}(\zeta) := \max_{\mu' \in \Delta(\mathcal{A})} f(\mu', \nu) - \min_{\nu' \in \Delta(\mathcal{B})} f(\mu, \nu') \le \varepsilon.$$

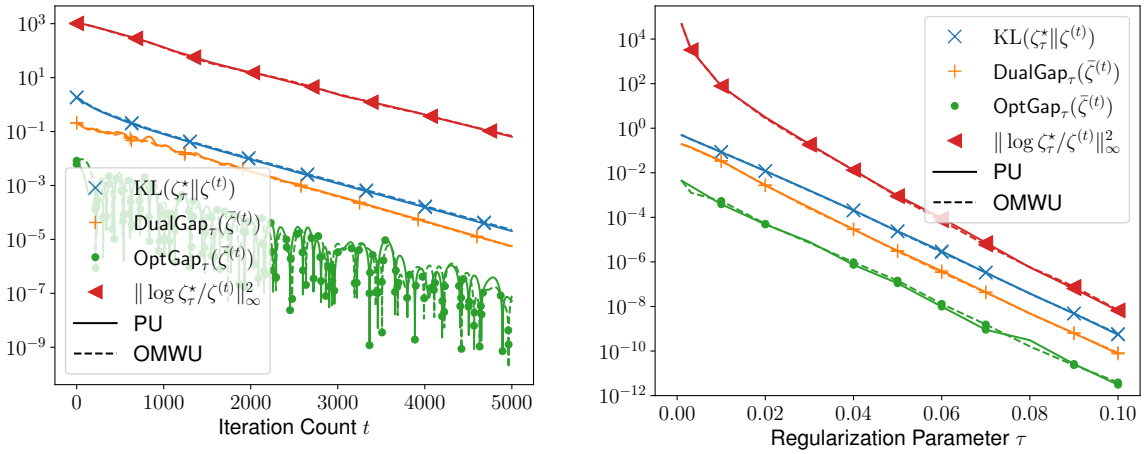


Figure 4.1: Performance illustration of the PU and OMWU methods for solving entropy-regularized matrix games with $|\mathcal{A}| = |\mathcal{B}| = 100$, where the entries of the payoff matrix A is generated independently from the uniform distribution on $[-1, 1]$. The learning rates are fixed as $\eta = 0.1$. The left panel plots various error metrics of convergence w.r.t. the iteration count with the entropy regularization parameter $\tau = 0.01$, while the right panel plots these error metrics at 1000-th iteration with different choices of τ. Due to their similar nature, PU and OMWU yield almost identical convergence behaviors and overlapping plots.

Observe that setting $\tau = \frac{\varepsilon/4}{\log|\mathcal{A}|+\log|\mathcal{B}|}$ guarantees

$$|f_\tau(\mu,\nu) - f(\mu,\nu)| < \varepsilon/4 \qquad \text{for all} \quad (\mu,\nu) \in \mathcal{A}\times\mathcal{B}$$

in view of the boundedness of the Shannon entropy $\mathcal{H}(\cdot)$. Theorem 4.9 (cf. (4.9d)) also ensures that our proposed algorithms find an approximate QRE $\bar{\zeta}^{(T)}$ such that $\mathsf{DualGap}_\tau(\bar{\zeta}^{(T)}) \le \varepsilon/2$ after taking $T = \widetilde{\mathcal{O}}\left(\frac{1}{\eta\varepsilon}\right)$ iterations, which is no more than

$$\widetilde{\mathcal{O}}\left(\frac{\|A\|_\infty}{\varepsilon}\right)$$

iterations with optimized learning rates. It follows immediately that

$$\mathsf{DualGap}(\bar{\zeta}^{(T)}) \le \mathsf{DualGap}_\tau(\bar{\zeta}^{(T)}) + \max_{\mu',\nu'}\left|f_\tau(\mu',\bar{\nu}^{(T)}) - f_\tau(\bar{\mu}^{(T)},\nu') - (f(\mu',\bar{\nu}^{(T)}) - f(\bar{\mu}^{(T)},\nu'))\right| \le \varepsilon, \tag{4.10}$$

and therefore $\bar{\zeta}^{(T)}$ is an ε-NE. Intriguingly, unlike prior work [Daskalakis and Panageas, 2019, Wei et al., 2021a] that analyzed the last-iterate convergence of OMWU in the unregularized setting ($\tau = 0$), our last-iterate convergence does not require the NE of (4.1) to be unique. See Table 1.2 for further comparisons.

Remark 6. *For simplicity, we have set the regularization parameter τ on the order of the final accuracy ε. In practice, it might be desirable to use an annealing schedule of τ similar to the doubling trick, see e.g. Yang et al. [2020], Li et al. [2021]. We omit such straightforward generalizations for conciseness.*

Rationality. Another attractive feature of the algorithms developed above is being *rational* (as introduced in Bowling and Veloso [2001]) in the sense that the algorithm returns the best-response policy of one player when the opponent takes any *fixed* stationary policy. More specially, in terms of matrix games, when player 2 sticks to a stationary policy ν, the update of player 1 reduces to

$$\mu^{(t+1)}(a) \propto \mu^{(t)}(a)^{1-\eta\tau} \exp(\eta[A\nu]_a). \tag{4.11}$$

In this case, Theorem 5 can be established in exactly the same fashion by restricting attention only to the updates of $\mu^{(t)}$.

4.4 Discussion

This work develops provably efficient policy extragradient methods (PU and OMWU) for entropy-regularized matrix games, whose last iterates are guaranteed to converge linearly to the quantal response equilibrium at a linear rate. Encouragingly, the rate of convergence is independent of the dimension of the problem, i.e. the sizes of the space space and the action space. In addition, the last iterates of the proposed algorithms can also be used to locate Nash equilibria for the unregularized competitive games without assuming the uniqueness of the Nash equilibria by judiciously tuning the amount of regularization.

This work opens up interesting opportunities for further investigations of policy extragradient methods for solving competitive games. For example, can we develop a two-time-scale policy extragradient algorithms for Markov games where the Q-function is updated simultaneously with the policy but potentially at a different time scale, using samples, such as in an actor-critic algorithm [Konda and Tsitsiklis, 2000]? This question is partially answered under exact gradient evaluation in Chapter 5. Can we generalize the proposed algorithms to handle more general regularization terms, similar to what has been accomplished in the single-agent setting [Lan, 2023, Zhan et al., 2023a]? Can we generalize the proposed algorithm to other type of games [Ao et al., 2023]? We leave the answers to future work.

Chapter 5

Two-player Zero-sum Markov Games

In this chapter, we formulate the problem of two-player zero-sum Markov game. We present a noval policy optimization method along with its theoretical guarantees. For more details and entire analysis, please refer to Cen et al. [2023].

5.1 Algorithm and theory: the infinite-horizon setting

5.1.1 Problem formulation

Two-player zero-sum discounted Markov game. A two-player zero-sum discounted Markov game is defined by a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{B}, P, r, \gamma)$, with finite state space $\mathcal{S}$, finite action spaces of the two players $\mathcal{A}$ and $\mathcal{B}$, reward function $r : \mathcal{S} \times \mathcal{A} \times \mathcal{B} \to [0, 1]$, transition probability kernel $P : \mathcal{S} \times \mathcal{A} \times \mathcal{B} \to \Delta(\mathcal{S})$ and discount factor $0 \leq \gamma < 1$. The action selection rule of the max player (resp. the min player) is represented by $\mu : \mathcal{S} \to \Delta(\mathcal{A})$ (resp. $\nu : \mathcal{S} \to \Delta(\mathcal{B})$), where the probability of selecting action $a \in \mathcal{A}$ (resp. $b \in \mathcal{B}$) in state $s \in \mathcal{S}$ is specified by $\mu(a|s)$ (resp. $\nu(b|s)$). The probability of transitioning from state s to a new state s' upon selecting the action pair $(a, b) \in \mathcal{A}, \mathcal{B}$ is given by $P(s'|s, a, b)$.

Value function and Q-function. For a given policy pair μ, ν, the state value of $s \in \mathcal{S}$ is evaluated by the expected discounted sum of rewards with initial state $s_0 = s$:

$$\forall s \in \mathcal{S} : \qquad V^{\mu,\nu}(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t, b_t) \big| s_0 = s\right], \tag{5.1}$$

the quantity the max player seeks to maximize while the min player seeks to minimize. Here, the trajectory $(s_0, a_0, b_0, s_1, \cdots)$ is generated according to $a_t \sim \mu(\cdot|s_t)$, $b_t \sim \nu(\cdot|s_t)$ and $s_{t+1} \sim P(\cdot|s_t, a_t, b_t)$. Similarly, the Q-function $Q^{\mu,\nu}(s, a, b)$ evaluates the expected discounted cumulative reward with initial state s and initial action pair (a, b):

$$\forall (s, a, b) \in \mathcal{S} \times \mathcal{A} \times \mathcal{B} : \qquad Q^{\mu,\nu}(s, a, b) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t, b_t) \big| s_0 = s, a_0 = a, b_0 = b\right]. \tag{5.2}$$

For notation simplicity, we denote by $Q^{\mu,\nu}(s) \in \mathbb{R}^{|\mathcal{A}|\times|\mathcal{B}|}$ the matrix $[Q^{\mu,\nu}(s, a, b)]_{(a,b)\in\mathcal{A}\times\mathcal{B}}$, so that

$$\forall s \in \mathcal{S} : \qquad V^{\mu,\nu}(s) = \mu(s)^\top Q^{\mu,\nu}(s)\nu(s).$$

Shapley [1953] proved the existence of a policy pair $(\mu^\star, \nu^\star)$ that solves the min-max problem

$$\max_{\mu} \min_{\nu} V^{\mu,\nu}(s)$$

for all $s \in \mathcal{S}$ simultaneously, and that the mini-max value is unique. A set of such optimal policy pair $(\mu^\star, \nu^\star)$ is called the Nash equilibrium (NE) to the Markov game.

Entropy regularized two-player zero-sum Markov game. Entropy regularization is shown to provably accelerate convergence in single-agent RL [Geist et al., 2019, Mei et al., 2020b, Cen et al., 2022b] and facilitate the analysis in two-player zero-sum matrix games [Cen et al., 2021] as well as Markov games [Cen et al., 2021, Zeng et al., 2022]. The entropy-regularized value function $V_\tau^{\mu,\nu}(s)$ is defined as

$$\forall s \in \mathcal{S}: \qquad V_\tau^{\mu,\nu}(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \Big(r(s_t, a_t, b_t) - \tau \log \mu(a_t|s_t) + \tau \log \nu(b_t|s_t)\Big) \Big| s_0 = s\right], \tag{5.3}$$

where $\tau \geq 0$ is the regularization parameter. Similarly, the regularized Q-function $Q_\tau^{\mu,\nu}$ is given by

$$\forall (s,a,b) \in \mathcal{S} \times \mathcal{A} \times \mathcal{B}: \qquad Q_\tau^{\mu,\nu}(s) = r(s,a,b) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a,b)}\left[V_\tau^{\mu,\nu}(s')\right]. \tag{5.4}$$

It is known that [Cen et al., 2021] there exists a unique pair of policy $(\mu_\tau^\star, \nu_\tau^\star)$ that solves the min-max entropy-regularized problem

$$\max_{\mu} \min_{\nu} \; V_\tau^{\mu,\nu}(s), \tag{5.5a}$$

or equivalently

$$\max_{\mu} \min_{\nu} \; \mu(s)^\top Q_\tau^{\mu,\nu}(s)\nu(s) + \tau \mathcal{H}\big(\mu(s)\big) - \tau \mathcal{H}\big(\nu(s)\big) \tag{5.5b}$$

for all $s \in \mathcal{S}$, and we call $(\mu_\tau^\star, \nu_\tau^\star)$ the quantal response equilibrium (QRE) [McKelvey and Palfrey, 1995] to the entropy-regularized Markov game. We denote the associated regularized value function and Q-function by

$$V_\tau^\star(s) = V_\tau^{\mu_\tau^\star, \nu_\tau^\star}(s) \quad \text{and} \quad Q_\tau^\star(s,a,b) = Q_\tau^{\mu_\tau^\star, \nu_\tau^\star}(s,a,b).$$

Goal. We seek to find an ε-optimal QRE or ε-QRE (resp. ε-optimal NE or ε-NE) $\zeta = (\mu, \nu)$ which satisfies

$$\max_{s \in \mathcal{S}, \mu', \nu'} \Big(V_\tau^{\mu',\nu}(s) - V_\tau^{\mu,\nu'}(s)\Big) \leq \varepsilon \tag{5.6}$$

(resp. $\max_{s \in \mathcal{S}, \mu', \nu'} \Big(V^{\mu',\nu}(s) - V^{\mu,\nu'}(s)\Big) \leq \varepsilon$) in a computationally efficient manner. In truth, the solution concept of ε-QRE provides an approximation of ε-NE with appropriate choice of the regularization parameter τ. Basic calculations tell us that

$$\begin{aligned} V^{\mu',\nu}(s) - V^{\mu,\nu'}(s) &= \big(V_\tau^{\mu',\nu}(s) - V_\tau^{\mu,\nu'}(s)\big) + \big(V^{\mu',\nu}(s) - V_\tau^{\mu',\nu}(s)\big) - \big(V^{\mu,\nu'}(s) - V_\tau^{\mu,\nu'}(s)\big) \\ &\leq V_\tau^{\mu',\nu}(s) - V_\tau^{\mu,\nu'}(s) + \frac{\tau(\log|\mathcal{A}| + \log|\mathcal{B}|)}{1-\gamma}, \end{aligned}$$

which guarantees that an $\varepsilon/2$-QRE is an ε-NE as long as $\tau \leq \frac{(1-\gamma)\varepsilon}{2(\log|\mathcal{A}|+\log|\mathcal{B}|)}$. For technical convenience, we assume

$$\tau \leq \frac{1}{\max\{1, \log|\mathcal{A}| + \log|\mathcal{B}|\}} \tag{5.7}$$

throughout the paper. In addition, one might instead be interested in the expected (entropy-regularized) value function when the initial state is sampled from a distribution $\rho \in \Delta(\mathcal{S})$ over $\mathcal{S}$, which are given by

$$V_\tau^{\mu,\nu}(\rho) := \mathop{\mathbb{E}}_{s\sim\rho}\left[V_\tau^{\mu,\nu}(s)\right], \qquad \text{and} \qquad V^{\mu,\nu}(\rho) := \mathop{\mathbb{E}}_{s\sim\rho}\left[V^{\mu,\nu}(s)\right].$$

The ε-QRE/NE can be defined analogously, which facilitates comparisons to a number of related works.

Additional notation. For notation convenience, we denote by ζ the concatenation of a policy pair μ and ν, i.e., $\zeta = (\mu, \nu)$. The QRE to the regularized problem is denoted by $\zeta_\tau^\star = (\mu_\tau^\star, \nu_\tau^\star)$. We use shorthand notation $\mu(s)$ and $\nu(s)$ to denote $\mu(\cdot|s)$ and $\nu(\cdot|s)$. In addition, we write $\mathsf{KL}\left(\mu(s) \,\|\, \mu'(s)\right)$ and $\mathsf{KL}\left(\nu(s) \,\|\, \nu'(s)\right)$ as $\mathsf{KL}_s\big(\mu \,\|\, \mu'\big)$ and $\mathsf{KL}_s\big(\nu \,\|\, \nu'\big)$, and let

$$\mathsf{KL}_s\big(\zeta \,\|\, \zeta'\big) = \mathsf{KL}_s\big(\mu \,\|\, \mu'\big) + \mathsf{KL}_s\big(\nu \,\|\, \nu'\big).$$

5.1.2 Single-loop algorithm design

In this section, we propose a single-loop policy optimization algorithm for finding the QRE of the entropy-regularized Markov game, which is generalized from the entropy-regularized OMWU method [Cen et al., 2021] for solving entropy-regularized matrix games, with a careful orchestrating of the policy update and the value update.

Review: entropy-regularized OMWU for two-player zero-sum matrix games. We briefly review the algorithm design of entropy-regularized OMWU method for two-player zero-sum matrix game [Cen et al., 2021], which our method builds upon. The problem of interest can be described as

$$\max_{\mu\in\Delta(\mathcal{A})} \min_{\nu\in\Delta(\mathcal{B})} \mu^\top A\nu + \tau\mathcal{H}(\mu) - \tau\mathcal{H}(\nu), \tag{5.8}$$

where $A \in \mathbb{R}^{|\mathcal{A}|\times|\mathcal{B}|}$ is the payoff matrix of the game. The update rule of entropy-regularized OMWU with learning rate $\eta > 0$ is defined as follows: $\forall a \in \mathcal{A}, b \in \mathcal{B}$,

$$\begin{cases} \mu^{(t)}(a) \propto \mu^{(t-1)}(a)^{1-\eta\tau} \exp(\eta[A\bar{\nu}^{(t)}]_a) \\ \nu^{(t)}(b) \propto \nu^{(t-1)}(b)^{1-\eta\tau} \exp(-\eta[A^\top\bar{\mu}^{(t)}]_b) \end{cases}, \tag{5.9a}$$

$$\begin{cases} \bar{\mu}^{(t+1)}(a) \propto \mu^{(t)}(a)^{1-\eta\tau} \exp(\eta[A\bar{\nu}^{(t)}]_a) \\ \bar{\nu}^{(t+1)}(b) \propto \nu^{(t)}(b)^{1-\eta\tau} \exp(-\eta[A^\top\bar{\mu}^{(t)}]_b) \end{cases}. \tag{5.9b}$$

We remark that the update rule can be alternatively motivated from the perspective of natural policy gradient [Kakade, 2001, Cen et al., 2022b] or mirror descent [Lan, 2023, Zhan et al., 2023a] with optimistic updates. In particular, the midpoint $(\bar{\mu}^{(t+1)}, \bar{\nu}^{(t+1)})$ serves as a prediction of $(\mu^{(t+1)}, \nu^{(t+1)})$ by running one step of mirror descent. Cen et al. [2021] established that the last iterate of entropy-regularized OMWU converges to the QRE of the matrix game (5.8) at a linear rate $(1-\eta\tau)^t$, as long as the step size η is no larger than $\min\left\{\frac{1}{2\|A\|_\infty+2\tau}, \frac{1}{4\|A\|_\infty}\right\}$.

Single-loop algorithm for two-player zero-sum Markov games. In view of the similarity in the problem formulations of (5.5b) and (5.8), it is tempting to apply the aforementioned method to the Markov game in a state-wise manner, where the Q-function assumes the role of the payoff matrix. It is worth noting, however, that Q-function depends on the policy pair $\zeta = (\mu, \nu)$ and is hence changing concurrently with the update of the policy pair. We take inspiration from Wei et al. [2021b] and equip the entropy-regularized OMWU method with the following update rule that iteratively approximates the value function in an actor-critic fashion:

$$Q^{(t+1)}(s,a,b) = r(s,a,b) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a,b)} \left[V^{(t)}(s') \right],$$

where $V^{(t+1)}$ is updated as a convex combination of the previous $V^{(t)}$ and the regularized game value induced by $Q^{(t+1)}$ as well as the policy pair $\bar{\zeta}^{(t+1)} = (\bar{\mu}^{(t+1)}, \bar{\nu}^{(t+1)})$:

$$\begin{aligned} V^{(t+1)}(s) &= (1-\alpha_{t+1}) V^{(t)}(s) \\ &\quad + \alpha_{t+1} \big[\bar{\mu}^{(t+1)}(s)^\top Q^{(t+1)}(s) \bar{\nu}^{(t+1)}(s) + \tau \mathcal{H}\big(\bar{\mu}^{(t+1)}(s)\big) - \tau \mathcal{H}\big(\bar{\nu}^{(t+1)}(s)\big) \big]. \end{aligned} \tag{5.10}$$

The update of V becomes more conservative with a smaller learning rate α_t, hence stabilizing the update of policies. However, setting α_t too small slows down the convergence of V to $V_\tau^\star$. A key novelty—suggested by our analysis—is the choice of the constant learning rates $\alpha := \alpha_t = \eta\tau$ which updates at a slower timescale than the policy due to $\tau < 1$. This is in sharp contrast to the vanishing sequence $\alpha_t = \frac{2/(1-\gamma)+1}{2/(1-\gamma)+t}$ adopted in Wei et al. [2021b], which is essential in their analysis but inevitably leads to a much slower convergence. We summarize the detailed procedure in Algorithm 6. Last but not least, it is worth noting that the proposed method access the reward via "first-order information", i.e., either agent can only update its policy with the marginalized value function $Q(s)\nu(s)$ or $Q(s)^\top \mu(s)$. Update rules of this kind are instrumental in breaking the curse of multi-agents in the sample complexity when adopting sample-based estimates in (5.12), as we only need to estimate the marginalized Q-function rather than its full form [Li et al., 2022, Chen et al., 2021].

5.1.3 Theoretical guarantees

Below we present our main results concerning the last-iterate convergence of Algorithm 6 for solving entropy-regularized two-player zero-sum Markov games in the infinite-horizon discounted setting.

Theorem 6. *Setting $0 < \eta \le \frac{(1-\gamma)^3}{32000|\mathcal{S}|}$ and $\alpha_t = \eta\tau$, it holds for all $t \ge 0$ that*

$$\max \left\{ \frac{1}{|\mathcal{S}|} \sum_{s\in\mathcal{S}} \mathsf{KL}_s\big(\zeta_\tau^\star \,\|\, \zeta^{(t)}\big),\ \frac{1}{2|\mathcal{S}|} \sum_{s\in\mathcal{S}} \mathsf{KL}_s\big(\zeta_\tau^\star \,\|\, \bar{\zeta}^{(t)}\big),\ \frac{3\eta}{|\mathcal{S}|} \sum_{s\in\mathcal{S}} \big\| Q^{(t)}(s) - Q_\tau^\star(s) \big\|_\infty \right\}$$
$$\le \frac{3000}{(1-\gamma)^2\tau} \Big(1 - \frac{(1-\gamma)\eta\tau}{4}\Big)^t; \tag{5.13a}$$

and

$$\max_{s\in\mathcal{S},\mu,\nu} \Big(V_\tau^{\mu,\bar{\nu}^{(t)}}(s) - V_\tau^{\bar{\mu}^{(t)},\nu}(s) \Big) \le \frac{6000|\mathcal{S}|}{(1-\gamma)^3\tau} \max\left\{ \frac{8}{(1-\gamma)^2\tau}, \frac{1}{\eta} \right\} \Big(1 - \frac{(1-\gamma)\eta\tau}{4}\Big)^t. \tag{5.13b}$$

Theorem 6 demonstrates that as long as the learning rate η is small enough, the last iterate of Algorithm 6 converges at a linear rate for the entropy-regularized Markov game. Compared with prior literatures investigating on policy optimization, our analysis focuses on the last-iterate convergence of non-Euclidean updates in the presence of entropy regularization, which appears to be the first of its kind. Several remarks are in order, with detailed comparisons in Table 1.3.

Algorithm 6: Entropy-regularized OMWU for Discounted Two-player Zero-sum Markov Game

Input: Regularization parameter $\tau > 0$, learning rate for policy update $\eta > 0$, learning rate for value update $\{\alpha_t\}_{t=1}^{\infty}$.

Initialization: Set $\mu^{(0)}, \bar{\mu}^{(0)}$, $\nu^{(0)}$ and $\bar{\nu}^{(0)}$ as uniform policies; and set

$$Q^{(0)} = 0, \quad V^{(0)} = \tau(\log|\mathcal{A}| - \log|\mathcal{B}|).$$

for $t = 0, 1, \cdots$ **do**

  for *all* $s \in \mathcal{S}$ **do in parallel**

    When $t \geq 1$, update policy pair $\zeta^{(t)}(s)$ as:

$$\begin{cases} \mu^{(t)}(a|s) \propto \mu^{(t-1)}(a|s)^{1-\eta\tau} \exp(\eta[Q^{(t)}(s)\bar{\nu}^{(t)}(s)]_a) \\ \nu^{(t)}(b|s) \propto \nu^{(t-1)}(b|s)^{1-\eta\tau} \exp(-\eta[Q^{(t)}(s)^\top \bar{\mu}^{(t)}(s)]_b) \end{cases}. \tag{5.11a}$$

    Update policy pair $\bar{\zeta}^{(t+1)}(s)$ as:

$$\begin{cases} \bar{\mu}^{(t+1)}(a|s) \propto \mu^{(t)}(a|s)^{1-\eta\tau} \exp(\eta[Q^{(t)}(s)\bar{\nu}^{(t)}(s)]_a) \\ \bar{\nu}^{(t+1)}(b|s) \propto \nu^{(t)}(b|s)^{1-\eta\tau} \exp(-\eta[Q^{(t)}(s)^\top \bar{\mu}^{(t)}(s)]_b) \end{cases}. \tag{5.11b}$$

    Update $Q^{(t+1)}(s)$ and $V^{(t+1)}(s)$ as

$$\begin{cases} Q^{(t+1)}(s,a,b) &= r(s,a,b) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a,b)} \left[V^{(t)}(s')\right] \\ V^{(t+1)}(s) &= (1-\alpha_{t+1}) V^{(t)}(s) \\ &\quad + \alpha_{t+1}\left[\bar{\mu}^{(t+1)}(s)^\top Q^{(t+1)}(s) \bar{\nu}^{(t+1)}(s) + \tau\mathcal{H}\big(\bar{\mu}^{(t+1)}(s)\big) - \tau\mathcal{H}\big(\bar{\nu}^{(t+1)}(s)\big)\right] \end{cases}. \tag{5.12}$$

- **Linear convergence to the QRE.** Theorem 6 demonstrates that the last iterate of Algorithm 6 takes at most $\widetilde{\mathcal{O}}\left(\frac{1}{(1-\gamma)\eta\tau}\log\frac{1}{\varepsilon}\right)$ iterations to yield an ε-optimal policy in terms of the KL divergence to the QRE $\max_{s\in\mathcal{S}} \mathsf{KL}_s\big(\zeta_\tau^\star \,\|\, \bar{\zeta}^{(t)}\big) \leq \varepsilon$, the entrywise error of the regularized Q-function $\left\|Q^{(t)} - Q_\tau^\star\right\|_\infty \leq \varepsilon$, as well as the duality gap $\max_{s\in\mathcal{S},\mu,\nu}\left(V_\tau^{\mu,\bar{\nu}^{(t)}}(s) - V_\tau^{\bar{\mu}^{(t)},\nu}(s)\right) \leq \varepsilon$ at once. Minimizing the bound over the learning rate η, the proposed method is guaranteed to find an ε-QRE within

$$\widetilde{\mathcal{O}}\left(\frac{|\mathcal{S}|}{(1-\gamma)^4\tau}\log\frac{1}{\varepsilon}\right)$$

  iterations, which significantly improves upon the one-side convergence rate of Zeng et al. [2022].

- **Last-iterate convergence to ε-optimal NE.** By setting $\tau = \frac{(1-\gamma)\varepsilon}{2(\log|\mathcal{A}|+\log|\mathcal{B}|)}$, this immediately leads to provable last-iterate convergence to an ε-NE, with an iteration complexity of

$$\widetilde{\mathcal{O}}\left(\frac{|\mathcal{S}|}{(1-\gamma)^5\varepsilon}\right),$$

  which again outperforms the convergence rate of Wei et al. [2021b].

Remark 7. *The learning rate η is constrained to be inverse proportional to $|\mathcal{S}|$, which is for the worst case and can be potentially loosened for problems with a small concentrability coefficient.*

5.2 Algorithm and theory: the finite-horizon setting

Episodic two-player zero-sum Markov game. An episodic two-player zero-sum Markov game is defined by a tuple $\{\mathcal{S}, \mathcal{A}, \mathcal{B}, H, \{P_h\}_{h=1}^H, \{r_h\}_{h=1}^H\}$, with $\mathcal{S}$ being a finite state space, $\mathcal{A}$ and $\mathcal{B}$ denoting finite action spaces of the two players, and $H > 0$ the horizon length. Every step $h \in [H]$ admits a transition probability kernel $P_h : \mathcal{S}\times\mathcal{A} \to \Delta(\mathcal{S})$ and reward function $r_h : \mathcal{S}\times\mathcal{A}\times\mathcal{B} \to [0, 1]$. Furthermore, $\mu = \{\mu_h\}_{h=1}^H$ and $\{\nu_h\}_{h=1}^H$ denote the policies of the two players, where the probability of the max player choosing $a \in \mathcal{A}$ (resp. the min player choosing $b \in \mathcal{B}$) at time h is specified by $\mu_h(a|s)$ (resp. $\nu_h(a|s)$).

Entropy regularized value functions. The value function and Q-function characterize the expected cumulative reward starting from step h by following the policy pair μ, ν. For conciseness, we only present the definition of entropy-regularized value functions below and remark that the their un-regularized counterparts $V_h^{\mu,\nu}$ and $Q_h^{\mu,\nu}$ can be obtained by setting $\tau = 0$. We have

$$V_{h,\tau}^{\mu,\nu}(s) = \mathbb{E}\left[\sum_{h'=h}^{H} \left[r_{h'}(s_{h'}, a_{h'}, b_{h'}) - \tau \log \mu_{h'}(a_{h'}|s_{h'}) + \tau \log \nu_{h'}(b_{h'}|s_{h'})\right] \,\middle|\, s_h = s\right];$$
$$Q_{h,\tau}^{\mu,\nu}(s, a, b) = r_h(s, a, b) + \mathbb{E}_{s'\sim P_h(\cdot|s,a,b)}\left[V_{h+1,\tau}^{\mu,\nu}(s')\right].$$

The solution concept of NE and QRE are defined in a similar manner by focusing on the episodic versions of value functions. We again denote the unique QRE by $\zeta_\tau^\star = (\mu_\tau^\star, \nu_\tau^\star)$.

Proposed method and convergence guarantee It is straightforward to adapt Algorithm 6 to the episodic setting with minimal modifications, with detailed procedure showcased in Algorithm 7. The analysis, which substantially deviates from the discounted setting, exploits the structure of finite-horizon MDP and time-inhomogeneous policies, enabling a much larger range of learning rates as showed in the following theorem.

Theorem 7. *Setting $0 < \eta \le \frac{1}{8H}$ and $\alpha_t = \eta\tau$, it holds for all $h \in [H]$ and $t \ge T_h := (H - h)T_{\mathsf{start}}$ with $T_{\mathsf{start}} = \lceil \frac{1}{\eta\tau} \log H \rceil$ that*

$$\left\|Q_{h,\tau}^\star - Q_h^{(t)}\right\|_\infty \le (1 - \eta\tau)^{t-T_h} t^{H-h}; \tag{5.16a}$$

$$\max_{s\in\mathcal{S},\mu,\nu}\left(V_{h,\tau}^{\mu,\bar{\nu}^{(t)}}(s) - V_{h,\tau}^{\bar{\mu}^{(t)},\nu}(s)\right) \le 4(1 - \eta\tau)^{t-T_h} \max\left\{\frac{8H^2}{\tau}, \frac{1}{\eta}\right\}\left(\frac{8H}{\tau} + 6\eta t^{H-h+1}\right). \tag{5.16b}$$

Theorem 7 implies that the last iterate of Algorithm 7 takes no more than $\widetilde{\mathcal{O}}\big(HT_{\mathsf{start}} + \frac{H}{\eta\tau}\log\frac{1}{\varepsilon}\big) = \widetilde{\mathcal{O}}\big(\frac{H}{\eta\tau}\log\frac{1}{\varepsilon}\big)$ iterations for finding an ε-QRE. Minimizing the bound over the learning rate η, Algorithm 7 is guaranteed to find an ε-QRE in

$$\widetilde{\mathcal{O}}\left(\frac{H^2}{\tau}\log\frac{1}{\varepsilon}\right)$$

iterations, which translates into an iteration complexity of $\widetilde{\mathcal{O}}\left(\frac{H^3}{\varepsilon}\right)$ for finding an ε-NE in terms of the duality gap, i.e., $\max_{s\in\mathcal{S},h\in[H],\mu,\nu}\left(V_h^{\mu,\bar{\nu}^{(t)}}(s) - V_h^{\bar{\mu}^{(t)},\nu}(s)\right) \le \varepsilon$, by setting $\tau = \mathcal{O}\Big(\frac{\varepsilon}{H(\log|\mathcal{A}|+\log|\mathcal{B}|)}\Big)$.

Algorithm 7: Entropy-regularized OMWU for Episodic Two-player Zero-sum Markov Game

Input: Regularization parameter $\tau > 0$, learning rate for policy update $\eta > 0$, learning rate for value update $\{\alpha_t\}_{t=1}^{\infty}$.

Initialization: Set $\mu^{(0)}, \bar{\mu}^{(0)}$, $\nu^{(0)}$ and $\bar{\nu}^{(0)}$ as uniform policies; set

$$Q^{(0)} = 0, \quad V^{(0)} = \tau(\log|\mathcal{A}| - \log|\mathcal{B}|).$$

for $t = 0, 1, \cdots$ **do**

  for *all* $h \in [H]$, $s \in \mathcal{S}$ **do in parallel**

    When $t \geq 1$, update policy pair $\zeta_h^{(t)}(s)$ as:

$$\begin{cases} \mu_h^{(t)}(a|s) \propto \mu_h^{(t-1)}(a|s)^{1-\eta\tau} \exp(\eta[Q_h^{(t)}(s)\bar{\nu}_h^{(t)}(s)]_a) \\ \nu_h^{(t)}(b|s) \propto \nu_h^{(t-1)}(b|s)^{1-\eta\tau} \exp(-\eta[Q_h^{(t)}(s)^\top \bar{\mu}_h^{(t)}(s)]_b) \end{cases}. \tag{5.14a}$$

    Update policy pair $\bar{\zeta}_h^{(t+1)}(s)$ as:

$$\begin{cases} \bar{\mu}_h^{(t+1)}(a|s) \propto \mu_h^{(t)}(a|s)^{1-\eta\tau} \exp(\eta[Q_h^{(t)}(s)\bar{\nu}_h^{(t)}(s)]_a) \\ \bar{\nu}_h^{(t+1)}(b|s) \propto \nu_h^{(t)}(b|s)^{1-\eta\tau} \exp(-\eta[Q_h^{(t)}(s)^\top \bar{\mu}_h^{(t)}(s)]_b) \end{cases}. \tag{5.14b}$$

    Update $Q_h^{(t+1)}(s)$ and $V_h^{(t+1)}(s)$ as

$$\begin{cases} Q_h^{(t+1)}(s,a,b) &= r_h(s,a,b) + \mathbb{E}_{s' \sim P_h(\cdot|s,a,b)}\left[V_{h+1}^{(t)}(s')\right] \\ V_h^{(t+1)}(s) &= (1-\alpha_{t+1})V_h^{(t)}(s) \\ &\quad + \alpha_{t+1}\left[\bar{\mu}_h^{(t+1)}(s)^\top Q_h^{(t+1)}(s)\bar{\nu}_h^{(t+1)}(s) + \tau\mathcal{H}\big(\bar{\mu}_h^{(t+1)}(s)\big) - \tau\mathcal{H}\big(\bar{\nu}_h^{(t+1)}(s)\big)\right] \end{cases}. \tag{5.15}$$

5.3 Discussion

This work develops policy optimization methods for zero-sum Markov games that feature single-loop and symmetric updates with provable last-iterate convergence guarantees. Our approach yields better iteration complexities in both infinite-horizon and finite-horizon settings, by adopting entropy regularization and non-Euclidean policy update. Important future directions include investigating whether larger learning rates are possible without knowing problem-dependent information a priori, extending the framework to allow function approximation, and designing sample-efficient implementations of the proposed method. Last but not least, the introduction of entropy regularization requires each agent to reveal the entropy of their current policy to each other, which prevents the proposed method from being fully *decentralized*. Can we bypass this by dropping the entropy information in value learning? We leave the answers to future work.

Chapter 6

Multi-player Zero-sum Polymatrix Games

In this chapter, we formulate the problem of multi-player zero-sum polymatrix game. We present single-timescale OWMU and two-timescale OMWU methods along with their theoretical guarantees. For more details and entire analysis, please refer to Ao et al. [2023].

6.1 Preliminaries

We start by introducing the formulation of zero-sum polymatrix games as well as the solution concept of NE and QRE.

Polymatrix games. We start by defining the polymatrix game.

Definition 2 (Polymatrix game)**.** *Let $\mathcal{G} := \{(V, E), \{S_i\}_{i\in V}, \{A_{ij}\}_{(i,j)\in E}\}$ be an n-player polymatrix game, where each element in the tuple is defined as follows.*

- *An undirected graph (V, E), with $V = [n]$ denoting the set of players and E the set of edges;*
- *For each player $i \in V$, S_i represents its action set, which is assumed to be finite;*
- *For each edge $(i, j) \in E$, $A_{ij} \in \mathbb{R}^{|S_i|\times|S_j|}$ and $A_{ji} \in \mathbb{R}^{|S_j|\times|S_i|}$ represent the payoff matrices associated with player i and j, i.e., when player i and player j choose $s_i \in S_i$ and $s_j \in S_j$, the received payoffs are given by $A_{ij}(s_i, s_j)$, $A_{ji}(s_j, s_i)$, respectively.*

Utility function. Given the strategy profile $s = (s_1, \cdots, s_n) \in S = \prod_{i\in V} S_i$ taken by all players, the utility function $u_i : S \to \mathbb{R}$ of player i is given by

$$u_i(s) = \sum\nolimits_{j:(i,j)\in E} A_{ij}(s_i, s_j).$$

Suppose that player i adopts a mixed/stochastic strategy or policy, $\pi_i \in \Delta(S_i)$, where the probability of selecting $s_i \in S_i$ is specified by $\pi_i(s_i)$. With slight abuse of notation, we denote the expected utility of player i with a mixed strategy profile $\pi = (\pi_1, \cdots, \pi_n) \in \Delta(S)$ as

$$u_i(\pi) = \mathop{\mathbb{E}}_{s_i\sim\pi_i,\forall i\in V}[u_i(s)] = \sum_{j:(i,j)\in E} \pi_i^\top A_{ij}\pi_j. \tag{6.1}$$

It turns out to be convenient to treat π_i and π as vectors in $\mathbb{R}^{|S_i|}$ and $\mathbb{R}^{\sum_{i\in V}|S_i|}$ without ambiguity, and concatenate all payoff matrices associated with player i into

$$A_i = (A_{i1}, \cdots, A_{in}) \in \mathbb{R}^{|S_i| \times \sum_{j\in V}|S_j|}, \tag{6.2}$$

where A_{ij} is set to 0 whenever $(i,j) \notin E$. In particular, it follows that $A_{ii} = 0$ for all $i \in V$. With these notation in place, we can rewrite the expected utility function (6.1) as

$$u_i(\pi) = \pi_i^\top A_i \pi, \tag{6.3}$$

where $A_i\pi \in \mathbb{R}^{|S_i|}$ can be interpreted as the expected utility of the actions in S_i for player i. In addition, we denote the maximum entrywise absolute value of payoff by $\|A\|_\infty = \max_{i,j}\|A_{ij}\|_\infty = \max_i \|A_i\|_\infty$, and the maximum degree of the graph by $d_{\max} = \max_{i\in V}\deg_i$, where $\deg_i$ is the degree of player i. Moreover, we denote $S_{\max} = \max_i |S_i|$ as the maximum size of the action space over all players.

Zero-sum polymatrix games. The game $\mathcal{G}$ is a zero-sum polymatrix game if it holds that

$$\sum\nolimits_{i\in V} u_i(s) = 0, \qquad \forall\, s \in S. \tag{6.4}$$

This also immediately implies that for any strategy profile $\pi \in \Delta(S)$, it follows that $\sum_{i\in V} u_i(\pi) = 0$.

Nash equilibrium (NE). A mixed strategy profile $\pi^\star = (\pi_1^\star, \cdots, \pi_n^\star)$ is a Nash equilibrium (NE) when each player i cannot further increase its own utility function u_i by unilateral deviation, i.e., $u_i(\pi_i', \pi_{-i}^\star) \le u_i(\pi_i^\star, \pi_{-i}^\star)$, for all $i \in V,\ \pi_i' \in \Delta(S_i)$, where the existence is guaranteed by the work [Cai et al., 2016]. Here we denote the mixed strategies of all players other than i by π_{-i} and write $u_i(\pi_i, \pi_{-i}) = u_i(\pi)$. To measure how close a strategy $\pi \in \Delta(S)$ is to an NE, we introduce

$$\texttt{NE-Gap}(\pi) = \max_{i\in V}\left[\max_{\pi_i'\in\Delta(S_i)} u_i(\pi_i', \pi_{-i}) - u_i(\pi)\right],$$

which measures the largest possible gain in the expected utility when players deviate from its strategy unilaterally. A mixed strategy profile π is called an ε-approximate Nash equilibrium (ε-NE) when $\texttt{NE-Gap}(\pi) \le \varepsilon$, which ensures that $u_i(\pi_i', \pi_{-i}) \le u_i(\pi_i, \pi_{-i}) + \varepsilon$, for all $i \in V,\ \pi_i' \in \Delta(S_i)$.

Quantal response equilibrium (QRE). The quantal response equilibrium (QRE), proposed by McKelvey and Palfrey [1995], generalizes the classical notion of NE under uncertain payoffs or bounded rationality, while balancing exploration and exploitation. A mixed strategy profile $\pi_\tau^\star = (\pi_{1,\tau}^\star, \cdots, \pi_{n,\tau}^\star)$ is a QRE when each player assigns its probability of action according to the expected utility of every action in a Boltzmann fashion, i.e., for all $i \in V$,

$$\pi_{i,\tau}^\star(k) = \frac{\exp([A_i\pi_\tau^\star]_k/\tau)}{\sum_{k\in S_i}\exp([A_i\pi_\tau^\star]_k/\tau)}, \qquad k \in S_i, \tag{6.5}$$

where $\tau > 0$ is the regularization parameter or temperature. Equivalently, this amounts to maximizing an entropy-regularized utility of each player [Mertikopoulos and Sandholm, 2016], i.e., $u_{i,\tau}(\pi_i', \pi_{-i,\tau}^\star) \le u_{i,\tau}(\pi_{i,\tau}^\star, \pi_{-i,\tau}^\star)$ for all $i \in V,\ \pi_i' \in \Delta(S_i)$. Here, the entropy-regularized utility function $u_i : S \to \mathbb{R}$ of player i is given by

$$u_{i,\tau}(\pi) = u_i(\pi) + \tau\mathcal{H}(\pi_i), \tag{6.6}$$

where $\mathcal{H}(\pi_i) = -\pi_i^\top \log \pi_i$ denotes the Shannon entropy of π_i. In Leonardos et al. [2021], it is shown that a unique QRE exists in a zero-sum polymatrix game. Similarly, we can measure the proximity of a strategy π to a QRE by

$$\texttt{QRE-Gap}_\tau(\pi) = \max_{i \in V} \left[\max_{\pi_i' \in \Delta(S_i)} u_{i,\tau}(\pi_i', \pi_{-i}) - u_{i,\tau}(\pi) \right]. \tag{6.7}$$

A mixed strategy profile π is called an ε-QRE when $\texttt{QRE-Gap}_\tau(\pi) \leq \varepsilon$. According to the straightforward relationship

$$\begin{aligned}
\texttt{NE-Gap}(\pi) &= \max_{i \in V} \left[\max_{\pi_i' \in \Delta(S_i)} u_i(\pi_i', \pi_{-i}) - u_i(\pi) \right] \\
&= \max_{i \in V} \left[\max_{\pi_i' \in \Delta(S_i)} u_{i,\tau}(\pi_i', \pi_{-i}) - u_{i,\tau}(\pi) + \tau(\pi_i')^\top \log \pi_i' - \tau \pi_i^\top \log \pi_i \right] \\
&\leq \max_{i \in V} \left[\max_{\pi_i' \in \Delta(S_i)} u_{i,\tau}(\pi_i', \pi_{-i}) - u_{i,\tau}(\pi) \right] + \tau \log S_{\max} \\
&= \texttt{QRE-Gap}_\tau(\pi) + \tau \log S_{\max},
\end{aligned} \tag{6.8}$$

it follows immediately that we can link an $\varepsilon/2$-QRE to ε-NE by setting $\tau = \frac{\varepsilon}{2 \log S_{\max}}$. This facilitates the translation of convergence to the QRE to one regarding the NE by appropriately setting the regularization parameter τ.

6.2 Performance guarantees of single-timescale OMWU

In this section, we present and study the entropy-regularized OMWU method [Cen et al., 2021] for finding the QRE of zero-sum polymatrix games. Whilst the method is originally proposed for finding QRE in a two-player zero-sum game, the update rule naturally generalizes to the multi-player setting as

$$\pi_i^{(t+1)}(k) \propto \pi_i^{(t)}(k)^{1-\eta\tau} \exp(\eta [A_i \overline{\pi}^{(t+1)}]_k), \qquad \forall k \in S_i, \tag{6.9}$$

where $\eta > 0$ is the learning rate and $\overline{\pi}^{(t+1)}$ serves as a prediction for $\pi^{(t+1)}$ via an extrapolation step

$$\overline{\pi}_i^{(t+1)}(k) \propto \pi_i^{(t)}(k)^{1-\eta\tau} \exp(\eta [A_i \overline{\pi}^{(t)}]_k), \qquad \forall k \in S_i. \tag{6.10}$$

In the asynchronous setting, however, each agent i receives a delayed payoff vector $A_i \overline{\pi}^{(\kappa_i^{(t)})}$ instead of $A_i \overline{\pi}^{(t)}$ in the t-th iteration, where

$$\kappa_i^{(t)} = \max\{t - \gamma_i^{(t)}, 0\}, \tag{6.11}$$

with $\gamma_i^{(t)} \geq 0$ representing the length of delay. The detailed procedure is outlined in Algorithm 8 using the single-timescale rule (6.12) for extrapolation.

6.2.1 Performance guarantees without delays

We first present our theorem concerning the last-iterate convergence of single-timescale OMWU for finding the QRE in the synchronous setting, i.e. $\gamma_i^{(t)} = 0$ for all $i \in V$ and $t \geq 0$. For any $\pi, \pi' \in V$, let $\mathsf{KL}(\pi \,\|\, \pi') = \sum_{i \in V} \mathsf{KL}(\pi_i \,\|\, \pi_i')$.

Algorithm 8: Entropy-regularized OMWU, agent i

1: Initialize $\pi_i^{(0)} = \overline{\pi}_i^{(0)}$ as uniform distribution. Learning rates η, and $\overline{\eta}$ (optional).
2: **for** $t = 0, 1, 2, \ldots$ **do**
3: Receive payoff vector $A_i \overline{\pi}^{(\kappa_i^{(t)})}$.
4: When $t \geq 1$, update π_i according to

$$\pi_i^{(t)}(k) \propto \pi_i^{(t-1)}(k)^{1-\eta\tau} \exp(\eta[A_i \overline{\pi}^{(\kappa_i^{(t)})}]_k), \qquad \forall k \in S_i.$$

5: Update $\overline{\pi}_i$ according to the single-timescale rule

$$\overline{\pi}_i^{(t+1)}(k) \propto \pi_i^{(t)}(k)^{1-\eta\tau} \exp(\eta[A_i \overline{\pi}^{(\kappa_i^{(t)})}]_k), \qquad \forall k \in S_i. \tag{6.12}$$

or the two-timescale rule

$$\overline{\pi}_i^{(t+1)}(k) \propto \pi_i^{(t)}(k)^{1-\overline{\eta}\tau} \exp(\overline{\eta}[A_i \overline{\pi}^{(\kappa_i^{(t)})}]_k), \qquad \forall k \in S_i. \tag{6.13}$$

6: **end for**

Theorem 8 (Last-iterate convergence without delays)**.** *Suppose that the learning rate η of single-timescale OMWU in Algorithm 8 obeys*

$$0 < \eta \leq \min\left\{\frac{1}{2\tau}, \frac{1}{4d_{\mathsf{max}} \|A\|_\infty}\right\}, \tag{6.14}$$

then for any $T \geq 0$, the iterates $\pi^{(T)}$ and $\overline{\pi}^{(T)}$ converge at a linear rate according to

$$\mathsf{KL}\left(\pi_\tau^\star \,\|\, \pi^{(T)}\right) \leq (1-\eta\tau)^T \mathsf{KL}\left(\pi_\tau^\star \,\|\, \pi^{(0)}\right), \quad \mathsf{KL}\left(\pi_\tau^\star \,\|\, \overline{\pi}^{(T+1)}\right) \leq 2(1-\eta\tau)^T \mathsf{KL}\left(\pi_\tau^\star \,\|\, \pi^{(0)}\right). \tag{6.15a}$$

Furthermore, the QRE-gap also converges linearly according to

$$\texttt{QRE-Gap}_\tau(\overline{\pi}^{(T)}) \leq \left(\eta^{-1} + 2\tau^{-1} d_{\mathsf{max}}^2 \|A\|_\infty^2\right)(1-\eta\tau)^{T-1} \mathsf{KL}\left(\pi_\tau^\star \,\|\, \pi^{(0)}\right). \tag{6.15b}$$

Theorem 8 demonstrates that as long as the learning rate η is sufficiently small, the last iterate of single-timescale OMWU converges to the QRE at a linear rate. Compared with prior works for finding approximate equilibrium for zero-sum polymatrix games, our approach features a closed-form multiplicative update and a fast linear last-iterate convergence. Some remarks are in order.

- *Linear convergence to the QRE.* Theorem 8 implies an iteration complexity of $\widetilde{\mathcal{O}}\left(\frac{1}{\eta\tau}\log\frac{1}{\varepsilon}\right)$ for finding an ε-QRE in a last-iterate manner, which leads to an iteration complexity of

$$\widetilde{\mathcal{O}}\left(\left(\frac{d_{\mathsf{max}}\|A\|_\infty}{\tau} + 1\right)\log\frac{1}{\varepsilon}\right)$$

by optimizing the learning rate in (6.14). The result is especially appealing as it avoids direct dependency on the number of agents n as well as the size of action spaces (up to logarithmic factors), suggesting that learning in competitive multi-agent games can be made quite scalable as long as the interactions among the agents are sparse (so that the maximum degree of the graph d_{max} is much smaller than the number of agents n).

- *Last-iterate convergence to ε-NE.* By setting τ appropriately, we end up with an iteration complexity of

$$\widetilde{\mathcal{O}}\left(\frac{d_{\mathsf{max}}\left\|A\right\|_{\infty}}{\varepsilon}\right)$$

for achieving last-iterate convergence to an ε-NE (cf. (6.8)), which outperforms the best existing last-iterate rate of $\widetilde{\mathcal{O}}\left(n\left\|A\right\|_{\infty}/\varepsilon^{2}\right)$ from Leonardos et al. [2021] by at least a factor of $n/(d_{\mathsf{max}}\varepsilon)$.

Remark 8. *Our results trivially extend to the setting of weighted zero-sum polymatrix games [Leonardos et al., 2021], which amounts to adopting different learning rates $\{\eta_i\}_{i\in V}$ at each player. In this case, the iteration complexity becomes $\widetilde{\mathcal{O}}\left(\max_{i\in V}\frac{1}{\eta_i\tau}\log\frac{1}{\varepsilon}\right)$.*

6.2.2 Performance guarantees under random delays

We continue to examine single-timescale OMWU in the more challenging asynchronous setting. In particularly, we show that the last iterate of single-timescale OMWU continues to converge linearly to the QRE at a slower rate, as long as the delays satisfy some mild statistical assumptions given below.

Assumption 4 (Random delays)**.** *Assume that for all $i\in V$, $t\geq 0$, the delay $\gamma_i^{(t)}$ is independently generated and satisfies*

$$\mathbb{E}_{\gamma_i^{(t)}\geq\ell}\left[\gamma_i^{(t)}\right]:=\mathbb{E}\left[\gamma_i^{(t)}\;\middle|\;\gamma_i^{(t)}\geq\ell\right]\leq E(\ell),\qquad\forall\ell=0,1,\ldots. \tag{6.16}$$

Additionally, there exists some constant $\zeta>1$, such that $L\triangleq\sum_{\ell=0}^{\infty}\zeta^{\ell}E(\ell)<\infty$.

We remark that Assumption 4 is a rather mild condition that applies to typical delay distributions, such as the Poisson distribution [Zhang et al., 2020e], as well as distributions with bounded support [Recht et al., 2011, Liu et al., 2014, Assran et al., 2020]. Roughly speaking, Assumption 4 implies that the probability of the delay decays exponentially with its length given that $\sum_{\ell=0}^{\infty}\zeta^{\ell}\mathbb{E}_{\gamma_i^{(t)}\geq\ell}\left[\gamma_i^{(t)}\right]=\mathbb{E}\left[\frac{\zeta^{\gamma_i^{(t)}+1}-1}{\zeta-1}\gamma_i^{(t)}\right]$,[1] where ζ^{-1} approximately indicates the decay rate. We have the following theorem.

Theorem 9 (Last-iterate convergence with random delays)**.** *Under Assumption 4, suppose that the regulari-zation parameter $\tau<\min\{1,d_{\mathsf{max}}\left\|A\right\|_{\infty}\}$ and the learning rate η of single-timescale OMWU in Algorithm 8 obeys*

$$0<\eta\leq\min\left\{\frac{\tau}{24d_{\mathsf{max}}^{2}\left\|A\right\|_{\infty}^{2}(L+1)},\frac{\zeta-1}{\tau\zeta}\right\}, \tag{6.17}$$

then for any $T\geq 1$, the iterates $\pi^{(T)}$ and $\overline{\pi}^{(T)}$ converges to $\pi_\tau^\star$ at the rate

$$\mathbb{E}\left[\mathsf{KL}\left(\pi_\tau^\star\,\|\,\pi^{(T)}\right)\right]\leq(1-\eta\tau)^{T}\,\mathsf{KL}\left(\pi_\tau^\star\,\|\,\pi^{(0)}\right),\qquad\mathbb{E}\left[\mathsf{KL}\left(\pi_\tau^\star\,\|\,\overline{\pi}^{(T)}\right)\right]\leq 2(1-\eta\tau)^{T}\mathsf{KL}\left(\pi_\tau^\star\,\|\,\pi^{(0)}\right). \tag{6.18a}$$

Furthermore, the QRE-gap also converges linearly according to

$$\mathbb{E}\left[\texttt{QRE-Gap}_{\tau}(\overline{\pi}^{(T)})\right]\leq 4\eta^{-1}(1-\eta\tau)^{T}\mathsf{KL}\left(\pi_\tau^\star\,\|\,\pi^{(0)}\right). \tag{6.18b}$$

[1] This can be checked by exchanging the order of sum: for $l\geq 0$, we have $\sum_{l=0}^{\infty}\mathbb{E}_{\gamma_i^{(t)}\geq l}\left[\gamma_i^{(t)}\right]=\sum_{k=0}^{\infty}\sum_{l=0}^{k}\zeta^{l}kP(\gamma_i^{(t)}=k)=\sum_{k=0}^{\infty}\frac{\zeta^{k+1}-1}{\zeta-1}kP(\gamma_i^{(t)}=k)=\mathbb{E}\left[\frac{\zeta^{\gamma_i^{(t)}+1}-1}{\zeta-1}\gamma_i^{(t)}\right]$.

Theorem 9 suggests that the iteration complexity to ε-QRE is no more than

$$\widetilde{\mathcal{O}}\left(\max\left\{d_{\mathsf{max}}^2\left\|A\right\|_\infty^2(L+1), \frac{\zeta}{\zeta-1}\right\}\frac{1}{\tau^2}\log\frac{1}{\varepsilon}\right)$$

after optimizing the learning rate, whose range is more limited compared with the requirement in (6.14) without delays. In particular, the range of the learning rate is proportional to the regularization parameter τ, an issue we shall try to address by resorting to two-timescale learning rates in OMWU. To facilitate further understanding, we showcase the iteration complexity for finding ε-QRE/NE under two typical scenarios: bounded delay and Poisson delay.

- *Bounded random delay.* When the delays are bounded above by some maximum delay γ, Assumption 4 is met with $\zeta = 1+\gamma^{-1}$ and $L = e\gamma(\gamma+1)$. Plugging into Theorem 9 yields an iteration complexity of

  $$\widetilde{\mathcal{O}}\left(\frac{d_{\mathsf{max}}^2\left\|A\right\|_\infty^2(\gamma+1)^2}{\tau^2}\log\frac{1}{\varepsilon}\right)$$

  for finding an ε-QRE, or

  $$\widetilde{\mathcal{O}}\left(\frac{d_{\mathsf{max}}^2\left\|A\right\|_\infty^2(\gamma+1)^2}{\varepsilon^2}\right) \tag{6.19}$$

  for finding an ε-NE, which increases quadratically as the maximum delay increases. Note that these rates are worse than those without delays (cf. Theorem 8).

- *Poisson delay.* When the delays follow the Poisson distribution with parameter $1/\overline{T}$, it suffices to set $\zeta = 1+\overline{T}^{-1}$ and $L = e\overline{T}(1+\overline{T})$ Assumption 4. This leads to an iteration complexity of

  $$\widetilde{\mathcal{O}}\left(\frac{d_{\mathsf{max}}^2\left\|A\right\|_\infty^2\overline{T}^2}{\tau^2}\log\frac{1}{\varepsilon}\right)$$

  for finding an ε-QRE, or

  $$\widetilde{\mathcal{O}}\left(\frac{d_{\mathsf{max}}^2\left\|A\right\|_\infty^2\overline{T}^2}{\varepsilon^2}\right)$$

  for finding an ε-NE, which is similar to the bounded random delay case.

6.3 Performance guarantees of two-timescale OMWU

While Theorem 9 demonstrates provable convergence of single-timescale OMWU with random delays, it remains unclear whether the update rule can be better motivated in more general asynchronous settings, and whether the convergence can be further ensured under adversarial delays. Indeed, theoretical insights from previous literature [Mokhtari et al., 2020a, Cen et al., 2021] suggest the critical role of $\overline{\pi}^{(t)}$ as a predictive surrogate for $\pi^{(t)}$ in enabling fast convergence, which no longer holds when $\overline{\pi}^{(t)}$ is replaced by a delayed feedback from $\overline{\pi}^{(\kappa_i^{(t)})}$. To this end, we propose to replace the extrapolation update (6.12) with one equipped with a different learning rate:

$$\overline{\pi}_i^{(t+1)}(k) \propto \pi_i^{(t)}(k)^{1-\overline{\eta}\tau}\exp(\overline{\eta}[A_i\overline{\pi}^{(\kappa_i^{(t)})}]_k), \qquad \forall k\in S_i, \tag{6.20}$$

which adopts a larger learning rate $\overline{\eta} > \eta$ to counteract the delay. Intuitively, a choice of $\overline{\eta} \approx (\gamma_i^{(t)}+1)\eta$ would allow $\overline{\pi}^{(\kappa_i^{(t)})}$ to approximate $\pi^{(t)}$ by taking the intermediate updates $\{\pi^{(l)} : \kappa_i^{(t)} \le l < t\}$ into consideration. We refer to this update rule as the two-timescale entropy-regularized OMWU, whose detailed procedure is again outlined in Algorithm 8 using (6.13) for extrapolation.

6.3.1 Performance guarantees under constant and known delays

To highlight the potential benefit of learning rate separation, we start by studying the convergence of two-timescale OMWU in the asynchronous setting with constant and known delays, which has been studied in [Weinberger and Ordentlich, 2002, Flaspohler et al., 2021, Meng et al., 2023]. We have the following theorem, which reveals a a faster linear convergence to the QRE by using a delay-aware two-timescale learning rate design.

Theorem 10 (Last-iterate convergence with fixed delays)**.** *Suppose that the delays $\gamma_i^{(t)} = \gamma$ are fixed and known. Suppose that the learning rate η of two-timescale OMWU in Algorithm 8 satisfies*

$$\eta \leq \min\left\{\frac{1}{2\tau(\gamma+1)}, \frac{1}{5d_{\max}\left\|A\right\|_{\infty}(\gamma+1)^2}\right\}$$

and $\overline{\eta}$ is determined by $1-\overline{\eta}\tau = (1-\eta\tau)^{(\gamma+1)}$, then the last iterate $\pi^{(T)}$ and $\overline{\pi}^{(T)}$ converge to the QRE at a linear rate: for $T \geq \gamma$,

$$\max\left\{\mathsf{KL}\left(\pi_\tau^\star \,\|\, \pi^{(T+1)}\right), \frac{1}{2}\mathsf{KL}\left(\pi_\tau^\star \,\|\, \overline{\pi}^{(T-\gamma+1)}\right)\right\} \leq (1-\eta\tau)^{T+1}\mathsf{KL}\left(\pi_\tau^\star \,\|\, \pi^{(0)}\right) + (1-\eta\tau)^{T+1-\gamma}.$$

In addition, the QRE-gap converges linearly according to

$$\texttt{QRE-Gap}_\tau(\overline{\pi}^{(T-\gamma+1)}) \leq 2\max\left\{\frac{d_{\max}^2\left\|A\right\|_{\infty}^2}{\tau}, \frac{1}{\eta}\right\}\left((1-\eta\tau)^{T+1}\mathsf{KL}\left(\pi_\tau^\star \,\|\, \pi^{(0)}\right) + (1-\eta\tau)^{T+1-\gamma}\right).$$

By optimizing the learning rate η, Theorem 10 implies that two-timescale OMWU takes at most

$$\widetilde{\mathcal{O}}\left(\frac{d_{\max}\left\|A\right\|_{\infty}(\gamma+1)^2}{\tau}\log\frac{1}{\varepsilon}\right)$$

iterations to find an ε-QRE in a last-iterate manner, which translates to an iteration complexity of

$$\widetilde{\mathcal{O}}\left(\frac{d_{\max}\left\|A\right\|_{\infty}(\gamma+1)^2}{\varepsilon}\right)$$

for finding an ε-NE. This significantly improves over the iteration complexity of $\widetilde{\mathcal{O}}\big(d_{\max}^2\left\|A\right\|_{\infty}^2(\gamma+1)^2/\varepsilon^2\big)$ (cf. (6.19)) for single-timescale OMWU, verifying the positive role of adopting two-timescale learning rate in enabling faster convergence.

6.3.2 Performance guarantees with permuted bounded delays

The above result requires the exact information of the delay, which may not always be available. Motivated by the need to address arbitrary or even adversarial delays, we consider a more realistic scenario, where the payoff sequence arrives in a permuted order [Agarwal and Duchi, 2011] constrained by a maximum bounded delay [McMahan and Streeter, 2014, Wan et al., 2022].

Assumption 5 (Bounded delay)**.** *For any $i \in V$ and $t > 0$, it holds that $\gamma_i^{(t)} \leq \gamma$.*

Assumption 6 (Permuted feedback)**.** *For any $t > 0$, the payoff vector at the t-th iteration is received by agent i only once. The payoff at the 0-th iteration can be used multiple times.*

The following theorem unveils the convergence of two-timescale OMWU to the QRE in an average sense under permutated bounded delays.

Theorem 11 (Average-iterate convergence under permutated delays)**.** *Under Assumption 5 and 6, suppose that the learning rate η of two-timescale OMWU in Algorithm 8 satisfies*

$$\eta \le \min\left\{\frac{1}{2\tau(\gamma+1)}, \frac{1}{28 d_{\mathsf{max}} \|A\|_\infty (\gamma+1)^{5/2}}\right\},$$

and $\overline{\eta}$ is determined by $1-\overline{\eta}\tau = (1-\eta\tau)^{(\gamma+1)}$, then for $T > 2\gamma$, it holds that

$$\begin{aligned}
&\frac{1}{T-2\gamma}\max\left\{\sum_{t=2\gamma}^{T-1}\mathsf{KL}\left(\pi_\tau^\star \,\|\, \pi^{(t+1)}\right), \frac{1}{3}\sum_{t=2\gamma}^{T-1}\mathsf{KL}\left(\pi_\tau^\star \,\|\, \overline{\pi}^{(t-\gamma+1)}\right)\right\} \\
&\le \frac{1}{\eta\tau(T-2\gamma)}\left(\mathsf{KL}\left(\pi_{i,\tau}^\star \,\|\, \pi_i^{(0)}\right) + n\right) + \frac{24 n\gamma \log S_{\max}}{T-2\gamma}. \qquad (6.21)
\end{aligned}$$

Furthermore, the average QRE-gap can be bounded by

$$\begin{aligned}
&\frac{1}{T-2\gamma}\sum_{t=2\gamma}^{T-1} \texttt{QRE-Gap}_\tau(\overline{\pi}^{(t+1)}) \\
&\le \max\left\{\frac{3 d_{\mathsf{max}}^2 \|A\|_\infty^2}{2\tau}, \tau\right\}\left(\frac{1}{\eta\tau(T-2\gamma)}\left(\mathsf{KL}\left(\pi_{i,\tau}^\star \,\|\, \pi_i^{(0)}\right) + n\right) + \frac{36 n\gamma \log S_{\max}}{T-2\gamma}\right).
\end{aligned}$$

Theorem 11 guarantees that the best iterate among $\{\overline{\pi}^{(t)}\}_{2\gamma<t\le T}$ is an ε-QRE as long as T is on the order of

$$\widetilde{\mathcal{O}}\left(\frac{n d_{\mathsf{max}}^3 \|A\|_\infty^3 (\gamma+1)^{5/2}}{\tau^2\varepsilon}\right),$$

which translates to an iteration complexity of

$$\widetilde{\mathcal{O}}\left(\frac{n d_{\mathsf{max}}^3 \|A\|_\infty^3 (\gamma+1)^{5/2}}{\varepsilon^3}\right)$$

for finding an ε-NE. While the rate seems slower than the previous theorems, Theorem 11 holds under arguably the weakest delay assumptions, where it can be even adversarially bounded. We remark that the result in (6.21) also guarantees the convergence of the last iterate $\overline{\pi}^{(t)}$ to the QRE asymptotically, although without a finite-time rate. This is in sharp contrast to typical average-iterate analysis that only applies to $\frac{1}{T}\sum_{t=1}^T \overline{\pi}^{(t)}$ without implications on the convergence of the last iterate $\overline{\pi}^{(t)}$.

6.4 Discussion

This work studies asynchronous gradient play in zero-sum polymatrix games, by investigating the convergence behaviors of entropy-regularized OMWU with delayed feedbacks under two different schedules of the learning rates. We demonstrate that the single-timescale OMWU enjoys a linear last-iterate convergence to the QRE even under mild statistical delays, thus enables robust learning of gradient play. However, the presence of the delay limits the allowable range of learning rates and slows down the convergence. To mitigate the impact, we further show that the method benefits from adopting a two-timescale learning rate, by achieving a faster convergence when the delay is fixed and known and provable convergence in a more general setting with permuted feedback and bounded delay. This work leaves open a number of interesting questions:

- Can we further tighten the convergence analysis in terms of dependencies on salient parameters?
- Can we establish the last-iterate convergence of two-timescale OMWU under bounded delay?

Chapter 7

Multi-player Potential Games

In this chapter, we formulate the problem of multi-player potential game. We present independent NPG method along with its theoretical guarantees. For more details and entire analysis, please refer to Cen et al. [2022a].

7.1 Potential games with entropy regularization

In this section, we introduce the basics of potential games, as well as the incorporation of entropy regularization into its formulation.

7.1.1 Potential games

A strategic game $\mathcal{G} = \{N, \mathcal{A}, \{u_i\}_{i\in[N]}\}$ consists of N agents each with an individual utility or payoff function

$$u_i : \mathcal{A}^N \to [0, 1], \qquad i \in [N],$$

where $\mathcal{A}$ is, without loss of generality, a finite action space shared by all agents. The policy or mixed strategy of agent i is denoted by $\pi_i \in \Delta(\mathcal{A})$, which is a distribution over the action space $\mathcal{A}$. By an abuse of notation, let $u_i(\pi)$ denote agent i's expected utility function under the joint policy $\pi = \pi_1 \times \cdots \times \pi_N \in \Delta(\mathcal{A})^N$, i.e.,

$$u_i(\pi) = \mathbb{E}_{a_i\sim\pi_i, \forall i\in[N]} \left[u_i(\boldsymbol{a})\right],$$

where we denote the action profile $(a_1, \cdots, a_N)$ of all agents by $\boldsymbol{a} \in \mathcal{A}^N$. We shall often instead write $\boldsymbol{a} = (a_i, a_{-i})$ where $a_{-i} = \{a_j\}_{j\neq i}$ collects the actions of all agents but i; similarly, we write $\pi = (\pi_i, \pi_{-i})$, where $\pi_{-i} = \{\pi_j\}_{j\neq i}$ collects the policies of all agents but i.

The game $\mathcal{G}$ is said to be a potential game if there exists a potential function $\Phi : \mathcal{A}^N \to \mathbb{R}$ such that

$$u_i(a_i,\, a_{-i}) - u_i(a_i',\, a_{-i}) = \Phi(a_i,\, a_{-i}) - \Phi(a_i',\, a_{-i})$$

for any $a_i, a_i' \in \mathcal{A}$, $a_{-i} \in \mathcal{A}^{N-1}$ and $i \in [N]$. We assume that

$$0 \leq \Phi(\boldsymbol{a}) \leq \Phi_{\max}, \qquad \forall \boldsymbol{a} \in \mathcal{A}^N, \tag{7.1}$$

where $\Phi_{\max}$ upper bounds the potential function. An important special case of the potential game is when all the agents share the same utility function, known as the identical-interest game Monderer and Shapley [1996a]. It is straightforward to see that for an identical-interest game, we can set

$\Phi = u_i$ for all $i \in [N]$, and therefore $\Phi_{\max} = 1$ due to the fact that the individual payoff is bounded in $[0, 1]$.

By linearity of expectation, we have

$$u_i(\pi_i, \pi_{-i}) - u_i(\pi_i', \pi_{-i}) = \Phi(\pi_i, \pi_{-i}) - \Phi(\pi_i', \pi_{-i}),$$

where, again with slight abuse of notation, we denote

$$\Phi(\pi) = \mathbb{E}_{\boldsymbol{a}\sim\pi}\left[\Phi(\boldsymbol{a})\right] = \mathbb{E}_{a_i\sim\pi_i, \forall i\in[N]}\left[\Phi(\boldsymbol{a})\right],$$

for any $\pi_i, \pi_i' \in \Delta(\mathcal{A})$, $\pi_{-i} \in \Delta(\mathcal{A})^{N-1}$ and $i \in [N]$.

Nash equilibrium. We now introduce the important notion of *Nash equilibrium* in a potential game.

Definition 3 (Nash equilibrium)**.** *A joint policy $\pi^\star$ is called a Nash equilibrium (NE) when it holds that*

$$u_i(\pi_i, \pi_{-i}) \geq u_i(\pi_i', \pi_{-i}), \quad \forall \pi_i' \in \Delta(\mathcal{A}),\ \forall i \in [N].$$

In other words, every agent cannot improve its utility function by deviating from the current policy. It is known that there exists at least one NE in a strategic game with finite agents and actions Nash [1951]. It follows immediately that the policy or strategy profile maximizing Φ in a potential game is an NE.

Marginalized utility. Before continuing, let us introduce an important quantity called the marginalized utility $r_i^\pi : \mathcal{A} \to \mathbb{R}$:

$$r_i^\pi(a) = \mathbb{E}_{a_{-i}\sim\pi_{-i}}\left[u_i(a, a_{-i})\right], \tag{7.2}$$

which can be viewed as the "single-agent" payoff or reward function when the policies of other agents are fixed. It is immediate to see that the utility function u_i can be written as

$$u_i(\pi) = \mathbb{E}_{\boldsymbol{a}\sim\pi}\left[u_i(\boldsymbol{a})\right] = \mathbb{E}_{a\sim\pi_i}\left[r_i^\pi(a)\right] = \langle r_i^\pi, \pi_i\rangle.$$

Here and throughout this thesis, we shall often abuse the notation to treat π, π_i and $r_i^{(t)}$ as vectors.

7.1.2 Entropy-regularized potential games

The *quantal response equilibrium* (QRE) is proposed by McKelvey and Palfrey McKelvey and Palfrey [1995] as a seminal extension to the Nash equilibrium, which enables players to combat randomness in payoffs. A QRE or logit equilibrium $\pi_\tau^\star = \pi_{\tau,1}^\star \times \cdots \times \pi_{\tau,N}^\star$ necessitates every agent to maximize its own utility function with entropy regularization Mertikopoulos and Sandholm [2016], i.e.,

$$u_{i,\tau}(\pi_{\tau,i}^\star, \pi_{\tau,-i}^\star) \geq u_{i,\tau}(\pi_i', \pi_{\tau,-i}^\star), \quad \forall \pi_i' \in \Delta(\mathcal{A}),$$

where the entropy-regularized individual utility function is given by

$$u_{i,\tau}(\pi) = u_i(\pi) + \tau\mathcal{H}(\pi_i).$$

Here, $\pi = \pi_1 \times \cdots \times \pi_N$, $\tau > 0$ is the regularization parameter, and $\mathcal{H}(\pi_i) = -\sum_{a\in\mathcal{A}} \pi_i(a|s) \log \pi_i(a|s)$ is the Shannon entropy of the policy π employed by agent i. By introducing the regularized potential function

$$\Phi_\tau(\pi) = \Phi(\pi) + \tau\mathcal{H}(\pi) := \Phi(\pi) + \tau \sum_{i\in[N]} \mathcal{H}(\pi_i),$$

it is easy to verify

$$u_{i,\tau}(\pi_i,\, \pi_{-i}) - u_{i,\tau}(\pi_i',\, \pi_{-i}) = \Phi_\tau(\pi_i,\, \pi_{-i}) - \Phi_\tau(\pi_i',\, \pi_{-i}).$$

for any $\pi_i, \pi_i' \in \Delta(\mathcal{A}), \pi_{-i} \in \Delta(\mathcal{A})^{N-1}$ and $i \in [N]$, as long as the unregularized game is a potential game.

Fixed-point characterization of QRE. An equivalent interpretation of QRE is to let each agent assign the probability mass in its policy according to every action's utility in a bounded rationality fashion Selten [1989]:

$$\pi_{\tau,i}^\star(a) \propto \exp\left(r_i^{\pi_\tau^\star}(a)/\tau\right), \quad \forall i \in [N], \tag{7.3}$$

where $r_i^{\pi_\tau^\star}$ is the marginalized utility of $\pi_\tau^\star$ defined in (7.2). Note that the above relation defines a fixed-point equation of $\pi_\tau^\star$.

7.2 Finite-time global convergence of independent natural policy gradient methods

A popular approach in the game theory literature to find an NE of a potential game is for each agent to switch to the best or better response policy, one at a time, and is generally referred to as *best-response dynamics*. This approach converges to an NE in finite iterations Monderer and Shapley [1996b] and underlies the algorithm design of a considerable number of works on, e.g., cut games Christodoulou et al. [2006], congestion games Chien and Sinclair [2011], weakly acyclic games Young [2004], and, more recently, their extensions in the Markovian setting Song et al. [2022], Arslan and Yüksel [2016]. It is noted, however, that this approach isolates itself from the independent learning paradigm as the update sequence needs to be scheduled in a centralized manner that is not often possible. Therefore, it is greatly desirable to design independent update rules, where each agent updates simultaneously without observing the payoffs of other agents, that achieves faster convergence. In this section, we answer this call by developing the independent natural policy gradient method to solve (entropy-regularized) potential games with finite-time global convergence guarantees.

7.2.1 Independent natural policy gradient method

We consider the standard softmax parameterization, where every agent i generates its own policy π_{θ_i} parameterized with $\theta_i \in \mathbb{R}^{|\mathcal{A}|}$ through the softmax transform:

$$\pi_{\theta_i}(a) = \frac{\exp(\theta_i(a))}{\sum_{a\in\mathcal{A}} \exp(\theta_i(a))}.$$

Every agent i evaluates and updates its policy independently using the *natural policy gradient* (NPG) method Kakade [2001]:

$$\theta_i \leftarrow \theta_i + \eta(\mathcal{F}^{\theta_i})^\dagger \nabla_{\theta_i} u_{i,\tau}(\pi), \tag{7.4}$$

where $(\mathcal{F}^{\theta_i})^\dagger$ denotes the Moore-Penrose pseudo-inverse of the Fisher information matrix $\mathcal{F}^{\theta_i}$, which is defined as

$$\mathcal{F}^{\theta_i} = \mathbb{E}_{a\sim\pi_{\theta_i}(\cdot)}\left[(\nabla_{\theta_i} \log \pi_{\theta_i}(a))(\nabla_{\theta_i} \log \pi_{\theta_i}(a))^\top\right],$$

and $\eta > 0$ is the learning rate. Moreover, the gradient $\nabla_{\theta_i} u_{i,\tau}(\pi)$ can be expressed as

$$\nabla_{\theta_i} u_{i,\tau}(\pi) = r_i^{\pi} - \tau \log \pi_i - \tau \mathbf{1}.$$

It turns out that with some algebra, the NPG update rule (7.4) can be equivalently rewritten with respect to the policies in use Cen et al. [2022b]:

$$\pi_i^{(t+1)}(a) \propto \pi_i^{(t)}(a)^{1-\eta\tau} \exp(\eta r_i^{(t)}(a)), \tag{7.5}$$

where $\pi_i^{(t)}$ denotes agent i's policy in the t-th iteration, and $r_i^{(t)} := r_i^{\pi^{(t)}}$ denotes the marginalized utility of $\pi^{(t)}$ (cf. (7.2)). The complete procedure is summarized in Algorithm 9.

Algorithm 9: Independent NPG for Entropy-regularized Potential Games

Input: Regularization parameter $\tau > 0$, step size for policy update $\eta > 0$.
Initialization: Set $\pi_i^{(0)}$ as uniform policy for all $i \in [N]$.
for $t = 0, 1, \cdots$ **do**
  for *all agent* $i \in [N]$ **do in parallel**
    Observe agent i's marginalized utility $r_i^{(t)}$.
    Perform policy update

$$\pi_i^{(t+1)}(a) \propto \pi_i^{(t)}(a)^{1-\eta\tau} \exp(\eta r_i^{(t)}(a)).$$

To better understand the update rule (7.5) as well as prepare for follow-up analysis, we introduce $\pi_i^{\star(t)}$ to denote agent i's best-response policy in the t-th iteration, which is the policy that obeys

$$u_{i,\tau}(\pi_i^{\star(t)}, \pi_{-i}^{(t)}) = \max_{\pi_i'} u_{i,\tau}(\pi_i', \pi_{-i}^{(t)}). \tag{7.6}$$

It is easily seen that

$$\pi_i^{\star(t)}(a) \propto \exp(r_i^{(t)}(a)/\tau). \tag{7.7}$$

Therefore, the updated policy in (7.5) can be regarded as a multiplicative combination of the current policy $\pi_i^{(t)}$ and the best-response policy $\pi_i^{\star(t)}$, where the weight is controlled by the learning rate η. Note that the unregularized counterpart of the method is equivalent to Multiplicative Weights Update method (MWU) Littlestone and Warmuth [1994], Arora et al. [2012] or Hedge Freund and Schapire [1999].

7.2.2 Finite-time global convergence

We are now ready to present our main theorem concerning the finite-time global convergence of independent NPG for solving entropy-regularized potential games. We introduce

$$\texttt{NE-gap}(\pi) = \max_{i \in [N], \pi_i' \in \Delta(\mathcal{A})} \left[u_i(\pi_i', \pi_{-i}) - u_i(\pi_i, \pi_{-i})\right]$$

and

$$\texttt{QRE-gap}_\tau(\pi) = \max_{i \in [N], \pi_i' \in \Delta(\mathcal{A})} \left[u_{i,\tau}(\pi_i', \pi_{-i}) - u_{i,\tau}(\pi_i, \pi_{-i})\right]$$

to characterize how close the joint policy π is to an equilibrium. A joint policy π is said to be an ε-QRE (resp. ε-NE) when $\texttt{QRE-gap}(\pi) \le \varepsilon$ (resp. $\texttt{NE-gap}(\pi) \le \varepsilon$). For notational simplicity, we denote

$$\Phi_\tau^{(t)} := \Phi_\tau(\pi^{(t)}), \qquad \texttt{QRE-gap}_\tau^{(t)} := \texttt{QRE-gap}_\tau(\pi^{(t)}), \quad \text{and} \quad \texttt{NE-gap}^{(t)} := \texttt{NE-gap}(\pi^{(t)}).$$

Our main theorem is as follows.

Theorem 12. *Suppose that the learning rate η satisfies $\eta \le \frac{1}{2(\min\{\sqrt{N}, 2\Phi_{\max}\}+\tau)}$, then for independent NPG updates (7.5), it holds that*

$$\frac{1}{T}\sum_{t=1}^{T} \texttt{QRE-gap}_\tau^{(t)} \le \frac{2}{\eta\tau T}\Big(\tau \left\|\log \pi^{(0)} - \log \pi^{\star(0)}\right\|_\infty + \sqrt{2\eta T(\Phi_\tau^{(T)} - \Phi_\tau^{(0)})}\Big).$$

Theorem 12 suggests that the average iterate of independent NPG converges to an ε-QRE at a sublinear rate when we initialize it via uniform policies, as indicated in the following corollary.

Corollary 1. *Assume the independent NPG method is initialized with uniform policies at all agents. Setting the learning rate $\eta = 1/(2(\min\{\sqrt{N}, 2\Phi_{\max}\} + \tau))$ and $\tau = \mathcal{O}(1)$, then independent NPG updates ensure that $\frac{1}{T}\sum_{t=1}^{T} \texttt{QRE-gap}_\tau^{(t)} \le \varepsilon$ with at most*

$$T = \mathcal{O}\left(\frac{\min\{\sqrt{N}, \Phi_{\max}\}\Phi_{\max}}{\tau^2\varepsilon^2}\right)$$

iterations.

Finding approximate NEs. It is possible to leverage the entropy-regularized potential game to find an approximate NE by setting the regularization parameter sufficiently small. Note that

$$\begin{aligned}
\texttt{NE-gap}(\pi) &= \max_{i\in[N], \pi_i' \in \Delta(\mathcal{A})} \left[u_i(\pi_i', \pi_{-i}) - u_i(\pi_i, \pi_{-i})\right] \\
&\le \max_{i\in[N], \pi_i' \in \Delta(\mathcal{A})} \left[u_{i,\tau}(\pi_i', \pi_{-i}) - u_{i,\tau}(\pi_i, \pi_{-i})\right] + \max_{i\in[N], \pi_i' \in \Delta(\mathcal{A})} \left[-\tau\mathcal{H}(\pi_i') + \tau\mathcal{H}(\pi_i)\right] \\
&\le \texttt{QRE-gap}_\tau(\pi) + \tau \log|\mathcal{A}|.
\end{aligned}$$

Therefore, by setting the entropy regularization at

$$\tau = \frac{\varepsilon}{2\log|\mathcal{A}|},$$

with at most

$$T = \widetilde{\mathcal{O}}\left(\frac{\min\{\sqrt{N}, \Phi_{\max}\}\Phi_{\max}}{\varepsilon^4}\right)$$

iterations, we can ensure $\frac{1}{T}\sum_{t=1}^{T} \texttt{NE-gap}^{(t)} \le \varepsilon$.

Comparisons with prior art. Importantly, our iteration complexities do not depend on the size of the action space (up to logarithmic factors), which is in sharp contrast to existing analyses of potential games using other policy gradient approaches, such as direct PG [Zhang et al., 2024a, Leonardos et al., 2022, Ding et al., 2022, Mao et al., 2022] and NPG with log-barrier regularization [Zhang et al., 2022b], where the iteration complexity scales as $\widetilde{\mathcal{O}}\left(N|\mathcal{A}|\Phi_{\max}/\varepsilon^2\right)$ to find an ε-approximate NE. In comparison, while our rate $\widetilde{\mathcal{O}}\left(\min\{\sqrt{N}, \Phi_{\max}\}\Phi_{\max}/\varepsilon^4\right)$ is worse in terms of ε, it is almost independent of the size $|\mathcal{A}|$ of the action space, as well as exhibits only a sublinear dependency with the number of agents N, thus can be beneficial for problems with large action spaces and a large number of agents. Furthermore, for the special case of identical-interest games [Monderer and Shapley, 1996a] where $\Phi_{\max} = 1$, the convergence rate of our method simplifies to

$$\widetilde{\mathcal{O}}\left(\frac{1}{\varepsilon^4}\right),$$

which leads to the first method that achieves a dimension-free iteration complexity (up to a logarithmic factor) for finding an ε-NE without imposing any isolation assumptions.

7.3 Discussion

This work studies independent NPG methods for entropy-regularized potential games and develops a sublinear rate of convergence to quantal response equilibrium, which is independent of the size of the action spaces up to logarithmic factors and grows only sublinearly with respect to the number of agents. In addition, the method achieves the first *dimension-free* convergence rate for the important special case of identical-interest games, where the rate is independent of both the size of the action space and the number of agents. The approach can also be used as a smoothing technique to find Nash equilibria by setting the regularization parameter sufficiently small, without imposing the isolation assumption as often required in prior works. This thesis leaves open a number of interesting questions:

- Can we tighten the convergence rate in terms of the dependencies on ε?
- Can we extend the analysis to establish finite-time global convergence for Markov potential games?

We leave the answers to future work.

Part III

Principled AI alignment

Chapter 8

Reinforcement Learning from Human Feedback

In this chapter, we formulate the problem of reinforcement learning from human feedback (RLHF). We present principled learning algorithms for both online and offline settings along with theoretical guarantees. For more details and entire analysis, please refer to Cen et al. [2024].

8.1 Preliminaries

In RLHF, a language model is described by a policy π, which generates an answer $y \in \mathcal{Y}$ given prompt $x \in \mathcal{X}$ according to the conditional probability distribution $\pi(\cdot|x)$. The standard RLHF process consists of four stages: supervised fine-tuning (SFT), preference data generation, reward modeling, and RL fine-tuning. In the SFT stage, a language model π_{sft} is obtained by fine-tuning a pre-trained LLM with supervised learning. The remaining stages continue training by leveraging the preference data, which we elaborate below.

Reward modeling from preference data. An oracle (e.g., a human labeler or a scoring model) evaluates the quality of two answers y_1 and y_2 given prompt x and reveals its preference. A widely used approach for modelling the probability of pairwise preferences is the Bradley–Terry model [Bradley and Terry, 1952]:

$$\mathbb{P}(y_1 \succ y_2|x) = \frac{\exp(r^\star(x, y_1))}{\exp(r^\star(x, y_1)) + \exp(r^\star(x, y_2))} = \sigma(r^\star(x, y_1) - r^\star(x, y_2)), \tag{8.1}$$

where $y_1 \succ y_2$ indicates that y_1 is preferred over y_2, $r^\star : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is the ground truth reward function, and $\sigma : \mathbb{R} \to (0, 1)$ is the logistic function. A preference data sample is denoted by a tuple (x, y_+, y_-), where y_+ (resp. y_-) is the preferred (resp. unpreferred) answer in the comparison.

Given a preference dataset $\mathcal{D} = \{(x^i, y_+^i, y_-^i)\}$ composed of independent samples, the reward function r can be estimated by maximum likelihood estimation (MLE):

$$r_{\mathsf{MLE}} = \arg\min_r \ \ell(r, \mathcal{D}), \tag{8.2}$$

where $\ell(r, \mathcal{D})$ is the negative log-likelihood of $\mathcal{D}$, given as

$$\ell(r, \mathcal{D}) \coloneqq -\sum_{(x^i, y_+^i, y_-^i) \in \mathcal{D}} \log \sigma(r(x^i, y_+^i) - r(x^i, y_-^i)). \tag{8.3}$$

RL fine-tuning. Given a reward model r, we seek to fine-tune the policy π to achieve an ideal balance between the expected reward and its distance from an initial policy π_{ref}, which is typically the same as π_{sft}. This is achieved by maximizing the KL-regularized value function $J(r, \pi)$, defined as

$$J(r, \pi) = \mathop{\mathbb{E}}_{x\sim\rho, y\sim\pi(\cdot|x)} [r(x, y)] - \beta \mathop{\mathbb{E}}_{x\sim\rho} [\mathsf{KL}\left(\pi(\cdot|x) \,\|\, \pi_{\text{ref}}(\cdot|x)\right)], \tag{8.4}$$

where $\mathsf{KL}\left(\pi_1 \,\|\, \pi_2\right)$ is the KL divergence from π_1 to π_2, and $\beta > 0$ is a regularization parameter. Consequently, the RL fine-tuned policy π_r with respect to the reward r satisfies

$$\pi_r \coloneqq \arg\max_{\pi} J(r, \pi), \tag{8.5}$$

which admits a closed-form solution [Rafailov et al., 2023], i.e.,

$$\forall(x \times y) \in \mathcal{X} \times \mathcal{Y}: \qquad \pi_r(y|x) = \frac{\pi_{\text{ref}}(y|x) \exp(r(x, y)/\beta)}{Z(r, x)}. \tag{8.6}$$

Here, $Z(r, x)$ is a normalization factor given by

$$Z(r, x) = \sum_{y'\in\mathcal{Y}} \pi_{\text{ref}}(y'|x) \exp(r(x, y')/\beta). \tag{8.7}$$

Direct preference optimization. The closed-form solution (8.6) allows us to write the reward function r in turn as

$$r(x, y) = \beta(\log \pi_r(y|x) - \log \pi_{\text{ref}}(y|x) + \log Z(r, x)). \tag{8.8}$$

Plugging the above equation into the reward MLE (8.2), we obtain the seminal formulation of direct preference optimization (DPO) over the policy space [Rafailov et al., 2023],

$$\pi_{\mathsf{DPO}} = \arg\min_{\pi} - \sum_{(x^i, y^i_+, y^i_-)\in\mathcal{D}} \log\sigma\left(\beta\left(\log\frac{\pi(y^i_+|x)}{\pi_{\text{ref}}(y^i_+|x)} - \log\frac{\pi(y^i_-|x)}{\pi_{\text{ref}}(y^i_-|x)}\right)\right), \tag{8.9}$$

which avoids explicitly learning the reward model.

8.2 Value-incentivized preference optimization

A major caveat of the standard RLHF framework concerns the lack of accounting for reward uncertainty, which is known to be indispensable in the success of standard RL paradigms in both online and offline settings [Cesa-Bianchi et al., 2017, Rashidinejad et al., 2022]. This motivates us to investigate a principled mechanism that be easily integrated into the RLHF pipeline, while bypassing the difficulties of explicit uncertainty estimation in LLMs.

8.2.1 General framework

In view of the sub-optimality of naive MLE for reward estimation [Cesa-Bianchi et al., 2017, Rashidinejad et al., 2022], and motivated by the effectiveness of reward-biased MLE in online RL [Kumar and Becker, 1982, Liu et al., 2020a, 2024a], we propose to regularize the reward estimate via

$$J^\star(r) = \max_{\pi} J(r, \pi), \tag{8.10}$$

which measures the resulting value function for the given reward if one acts according to its optimal policy. However, in RLHF, by the definition (8.1), the reward function $r^\star$ is only identifiable up to a prompt-dependent global shift. Specifically, letting $r_1(x,y) = r_2(x,y)+c(x)$ be two reward functions that only differ by a prompt-dependent shift $c(x)$, we have $r_1(x,y_1)-r_1(x,y_2) = r_2(x,y_1)-r_2(x,y_2)$, which leads to $J^\star(r_1) = J^\star(r_2) + \mathbb{E}_{x\sim\rho}[c(x)]$. To resolve this challenge, we introduce the following equivalent class of reward functions for the Bradley-Terry model to eliminate the shift ambiguity, which also has the calibration effect of centering the reward function while offering a regularization mechanism to incorporate additional policy preferences.

Assumption 7. *We assume that $r^\star \in \mathcal{R}$, where*

$$\mathcal{R} = \left\{ r : \mathop{\mathbb{E}}_{x\sim\rho, y\sim\pi_{cal}(\cdot|x)} [r(x,y)] = 0. \right\}. \tag{8.11}$$

Here, ρ is the prompt distribution and π_{cal} is a fixed calibration distribution independent of the algorithm.

The proposed regularized MLE of the Bradley-Terry model (8.2) appends a bias term to the negative likelihood

$$r_{\mathsf{VPO}} = \arg\min_{r\in\mathcal{R}} \{\ell(r,\mathcal{D}) - \mathsf{sign}\cdot\alpha\cdot J^\star(r)\}, \tag{8.12}$$

incentivizing the algorithm to favor (resp. avoid) reward models with higher value $J^\star(r)$ in the online (resp. offline) setting. Here, $\alpha > 0$ is a constant controlling the strength of regularization, and sign is set to 1 in the online setting and -1 in the offline setting.

At first glance, the objective function for VPO (8.12) does not immediately imply a computationally-efficient algorithm due to the presence of $J^\star(r)$. However, by exploiting the same closed-form solution for the optimal policy given the reward in (8.6), and the reward representation inferred from the policy via (8.8), we can explicitly express $J^\star(r)$ as

$$\begin{aligned}
J^\star(r) &= \mathop{\mathbb{E}}_{x\sim\rho, y\sim\pi_r(\cdot|x)} [r(x,y) - \beta(\log\pi_r(y|x) - \log\pi_{\mathrm{ref}}(y|x))] \\
&= \mathop{\mathbb{E}}_{x\sim\rho, y\sim\pi_r(\cdot|x)} [\log Z(r,x)] \\
&= \mathop{\mathbb{E}}_{x\sim\rho, y\sim\pi_{\mathrm{cal}}(\cdot|x)} [\log Z(r,x)] \\
&= \mathop{\mathbb{E}}_{x\sim\rho, y\sim\pi_{\mathrm{cal}}(\cdot|x)} [r(x,y) - \beta(\log\pi_r(y|x) - \log\pi_{\mathrm{ref}}(y|x))] \\
&= -\beta \mathop{\mathbb{E}}_{x\sim\rho, y\sim\pi_{\mathrm{cal}}(\cdot|x)} [\log\pi_r(y|x) - \log\pi_{\mathrm{ref}}(y|x)]\,,
\end{aligned} \tag{8.13}$$

where the second step follows because the bracketed term is independent of y (c.f. (8.6)) and the last step follows from (8.11) whenever $r \in \mathcal{R}$. Given this key ingredient, we can then rewrite (8.12)

to directly optimize the LLM policy, in a flavor similar to DPO, as

$$
\begin{aligned}
&\pi_{\mathsf{VPO}} \\
&= \operatorname*{argmin}_{\pi_r:\, r\in\mathcal{R}} \{\ell(r,\mathcal{D}) - \mathsf{sign}\cdot\alpha\cdot J^\star(r)\} \\
&= \operatorname*{argmin}_{\pi_r:\, r\in\mathcal{R}} \Big\{ -\sum_{(x^i,y_+^i,y_-^i)\in\mathcal{D}} \log\sigma\Big(\beta\log\frac{\pi_r(y_+^i|x^i)}{\pi_{\mathrm{ref}}(y_+^i|x^i)} - \beta\log\frac{\pi_r(y_-^i|x^i)}{\pi_{\mathrm{ref}}(y_-^i|x^i)}\Big) \\
&\qquad\qquad + \mathsf{sign}\cdot\alpha\beta \mathop{\mathbb{E}}_{x\sim\rho,y\sim\pi_{\mathrm{cal}}(\cdot|x)} [\log\pi_r(y|x) - \log\pi_{\mathrm{ref}}(y|x)] \Big\} \\
&= \arg\min_{\pi} \Big\{ -\sum_{(x^i,y_+^i,y_-^i)\in\mathcal{D}} \log\sigma\Big(\beta\log\frac{\pi(y_+^i|x^i)}{\pi_{\mathrm{ref}}(y_+^i|x^i)} - \beta\log\frac{\pi(y_-^i|x^i)}{\pi_{\mathrm{ref}}(y_-^i|x^i)}\Big) \\
&\qquad\qquad + \mathsf{sign}\cdot\alpha\beta \mathop{\mathbb{E}}_{x\sim\rho,y\sim\pi_{\mathrm{cal}}(\cdot|x)} [\log\pi(y|x) - \log\pi_{\mathrm{ref}}(y|x)] \Big\},
\end{aligned}
\tag{8.14}
$$

where we drop the constraint on $r \in \mathcal{R}$, since for any policy π there exists $r \in \mathcal{R}$ such that $\pi = \pi_r$.

Observing that the reference policy $\pi_{\mathrm{ref}}(y|x)$ in the last term of (8.14) $\mathop{\mathbb{E}}_{x\sim\rho,y\sim\pi_{\mathrm{cal}}(\cdot|x)} [\log\pi(y|x) - \log\pi_{\mathrm{ref}}(y|x)]$ does not impact the optimization solution, we can change it to $\mathop{\mathbb{E}}_{x\sim\rho,y\sim\pi_{\mathrm{cal}}(\cdot|x)} [\log\pi(y|x) - \log\pi_{\mathrm{cal}}(y|x)] \Big\} = -\mathop{\mathbb{E}}_{x\sim\rho} [\mathsf{KL}\left(\pi_{\mathrm{cal}}(\cdot|x) \,\|\, \pi(\cdot|x)\right)]$, which amounts to adding a KL regularization to the original DPO, and offers an interesting interpretation as pushing π against/towards π_{cal} in the online/offline settings respectively, unveiling the role of reward calibration in RLHF.

In what follows, we elaborate the development of VPO in both the online and offline settings with corresponding theoretical guarantees under linear function approximation.

8.2.2 Online RLHF: algorithm and theory

The online RLHF procedure extends training by performing reward learning and policy learning iteratively, with a growing preference dataset collected by using the current policy. We use $\pi^{(t)}$ to denote the policy used in the t-th iteration, where the superscript $^{(t)}$ indicates iteration t in the online setting. The t-th iteration of VPO for online RLHF consists of the following steps:

1. **New preference data generation.** We sample a new prompt $x^{(t)} \sim \rho$ and two answers $y_1^{(t)}, y_2^{(t)} \sim \pi^{(t)}(\cdot|x^{(t)})$, query the preference oracle and append $(x^{(t)}, y_+^{(t)}, y_-^{(t)})$ to the preference dataset.

2. **Reward learning.** We train a reward model with preference data $\mathcal{D}^{(t)} := \{(x^{(s)}, y_+^{(s)}, y_-^{(s)})\}_{s=1}^t$ by minimizing the regularized negative log-likelihood, i.e.,

$$
r^{(t+1)} = \arg\min_{r\in\mathcal{R}} \{\ell(r,\mathcal{D}^{(t)}) - \alpha\cdot J^\star(r)\}. \tag{8.15}
$$

3. **Policy learning.** This step trains the policy by solving the RL fine-tuning problem:

$$
\pi^{(t+1)} = \arg\max_{\pi} J(r^{(t+1)}, \pi). \tag{8.16}
$$

We summarize the detailed procedure in Algorithm 10.

Algorithm 10: VPO for online RLHF

initialization: $\pi^{(0)}$.

for $t = 0, 1, 2, \cdots$ **do**

Sample $x^{(t)} \sim \rho$, $y_1^{(t)}, y_2^{(t)} \sim \pi^{(t)}(\cdot|x^{(t)})$.

Obtain the preference between $(x^{(t)}, y_1^{(t)})$ and $(x^{(t)}, y_2^{(t)})$ from some oracle. Denote the comparison outcome by $(x^{(t)}, y_+^{(t)}, y_-^{(t)})$.

Update policy π as

$$\pi^{(t+1)} = \operatorname*{argmin}_{\pi} \Big\{ -\sum_{s=1}^{t} \log \sigma\Big(\beta \log \frac{\pi(y_+^{(s)}|x^{(s)})}{\pi_{\mathrm{ref}}(y_+^{(s)}|x^{(s)})} - \beta \log \frac{\pi(y_-^{(s)}|x^{(s)})}{\pi_{\mathrm{ref}}(y_-^{(s)}|x^{(s)})}\Big) + \alpha\beta \mathop{\mathbb{E}}_{x\sim\rho, y\sim\pi_{\mathrm{cal}}(\cdot|x)} [\log \pi(y|x) - \log \pi_{\mathrm{ref}}(y|x)] \Big\}. \tag{8.17}$$

Theoretical analysis. Encouragingly, VPO admits appealing theoretical guarantees under function approximation. For simplicity, we restrict attention to linear approximation of the reward model.

Assumption 8 (Linear Reward)**.** *We parameterize the reward model by*

$$r_\theta(x, y) = \langle \phi(x, y), \theta \rangle, \quad \forall (x, y) \in \mathcal{X} \times \mathcal{Y}, \tag{8.18}$$

where $\phi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^d$ is a fixed feature mapping and $\theta \in \mathbb{R}^d$ is the parameters. We assume that $\|\phi(x, y)\|_2 \le 1$ for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$, and that $r^\star(x, y) = \langle \phi(x, y), \theta^\star \rangle$ for some $\theta^\star$.

Under Assumption 7 and 8, it is sufficient to focus on $\theta \in \Theta$ where

$$\Theta = \Big\{ \theta \in \mathbb{R}^d : \mathop{\mathbb{E}}_{x\sim\rho, y\sim\pi_{\mathrm{cal}}(\cdot|x)} [\langle \phi(x, y), \theta \rangle] = 0 \Big\}. \tag{8.19}$$

The next theorem demonstrates that Algorithm 10 achieves $\widetilde{\mathcal{O}}(\sqrt{T})$ cumulative regret under mild assumptions. The proof is provided in Appendix G.1. The proof logic follows from that of [Liu et al., 2024a].

Theorem 13. *Under Assumptions 7 and 8, let $r_{\theta^{(t)}} \in \Theta$ denote the corresponding reward model for $\pi^{(t)}$. Assume that $\|\theta^\star\|_2 \le C$ and $\|\theta^{(t)}\|_2 \le C, \forall t \ge 0$ for some $C > 0$. Then with probability $1 - \delta$ we have*

$$\mathsf{Regret} := \sum_{t=1}^{T} \big[J^\star(r^\star) - J(r^\star, \pi^{(t)}) \big] \le \widetilde{\mathcal{O}}(\exp(2C + C/\beta)\sqrt{\kappa d T}),$$

with $\alpha = \dfrac{1}{\exp(2C + C/\beta)} \sqrt{\dfrac{T}{\kappa \min\{d \log T, T\}}}$ *and* $\kappa = \sup_{x,y} \dfrac{\pi_{cal}(y|x)}{\pi_{ref}(y|x)}$.

Theorem 13 shows that VPO achieves the same $\widetilde{\mathcal{O}}(\sqrt{T})$ regret for online RLHF as its counterparts in standard contextual bandits with scalar rewards and using UCB for exploration [Lattimore and Szepesvári, 2020].

Algorithm 11: VPO for offline RLHF

input: offline preference data $\mathcal{D}$ of size N.

Get policy $\widehat{\pi}$ by optimizing

$$\widehat{\pi} = \arg\min_{\pi} \Big\{ -\sum_{i=1}^{N} \log \sigma\Big(\beta \log \frac{\pi(y_+^i|x^i)}{\pi_{\mathrm{ref}}(y_+^i|x^i)} - \beta \log \frac{\pi(y_-^i|x^i)}{\pi_{\mathrm{ref}}(y_-^i|x^i)}\Big) - \alpha\beta \mathop{\mathbb{E}}_{x\sim\rho, y\sim\pi_{\mathrm{cal}}(\cdot|x)} [\log \pi(y|x) - \log \pi_{\mathrm{ref}}(y|x)] \Big\}.$$

Remark 9. *The boundedness condition on $\{\theta^{(t)}\}_{t\geq 0}$ can be potentially mitigated by incoporating into the algorithm a norm regularization term on θ. Further investigation is needed to determine its precise algorithmic form in (8.17).*

Remark 10. *The analysis naturally extends to allowing mini-batch samples of size M in every iteration, yielding an improved regret bound scaled by $1/\sqrt{M}$ and α scaled by $\sqrt{M}$.*

8.2.3 Offline RLHF: algorithm and theory

In offline RLHF, a fixed offline preference dataset is collected $\mathcal{D} := \{x^i, y_+^i, y_-^i\}_{i=1}^N$, where $x^i \sim \rho$, $y^i \sim \pi_{\mathsf{b}}(\cdot|x)$ are sampled from a behavior policy π_{b}, such as π_{sft} from SFT. The proposed VPO for offline RLHF consists of one pass through the reward and policy learning phases, i.e.,

$$\widehat{r} = \arg\min_{r\in\mathcal{R}} \{\ell(r, \mathcal{D}) + \alpha \cdot J^\star(r)\} \qquad \text{and} \qquad \widehat{\pi} = \arg\max_{\pi} J(\widehat{r}, \pi), \tag{8.20}$$

which discourages over-optimization of the reward function given the limited offline preference data. In the same vein as deriving (8.17), and by leveraging (8.13), we obtain the direct policy update rule:

$$\widehat{\pi} = \arg\min_{\pi} \Big\{ -\sum_{i=1}^{N} \log \sigma\Big(\beta \log \frac{\pi(y_+^i|x^i)}{\pi_{\mathrm{ref}}(y_+^i|x^i)} - \beta \log \frac{\pi(y_-^i|x^i)}{\pi_{\mathrm{ref}}(y_-^i|x^i)}\Big) - \alpha\beta \mathop{\mathbb{E}}_{x\sim\rho, y\sim\pi_{\mathrm{cal}}(\cdot|x)} [\log \pi(y|x) - \log \pi_{\mathrm{ref}}(y|x)] \Big\}. \tag{8.21}$$

We summarize the detailed procedure in Algorithm 11. When π_{cal} is set to π_{ref}, the regularization term becomes the KL divergence between π and π_{ref}, which is reminiscent of a popular choice in offline RL practice [Kumar et al., 2020]. Another heuristic choice is to set π_{cal} to the marginalized positive answer distribution from the dataset, i.e., $(x, y_+) \sim \mathcal{D}$, which leads to a similar objective in [Pal et al., 2024].

Saddle-point characterization and pessimism. We first illustrate that VPO indeed executes the principle of pessimism in a complementary manner to the standard approach of pessimism, which finds a policy that maximizes the worst-case value function over a confidence set. In particular, this strategy [Uehara and Sun, 2022] obtains a policy by solving

$$\widehat{\pi}_{\mathsf{LCB}} = \arg\max_{\pi} \min_{r\in\mathcal{R}_\delta} J(r, \pi) \tag{8.22}$$

where the confidence set $\mathcal{R}_\delta$ is typically set to $\{r : \ell(r, \mathcal{D}) \le \ell(r_{\mathsf{MLE}}, \mathcal{D}) + \delta\}$ or $\{r : \mathsf{dist}(r, r_{\mathsf{MLE}}) \le \delta\}$ for some $\delta > 0$ and s distance measure dist. Turning to VPO, note that by (8.20) we have

$$\widehat{r} = \arg\min_r \big\{\ell(r, \mathcal{D}) + \alpha J^\star(r)\big\} = \arg\min_r \max_\pi \big\{\ell(r, \mathcal{D}) + \alpha J(r, \pi)\big\}. \tag{8.23}$$

Since $\ell(r, \mathcal{D}) + \alpha J(r, \pi)$ is strongly concave over π, and convex over r, it allows us to formulate $(\widehat{r}, \widehat{\pi})$ as a saddle point in the following lemma. The proof is given in Appendix G.2.1.

Lemma 2. *$(\widehat{r}, \widehat{\pi})$ is a saddle point of the objective $\ell(r, \mathcal{D}) + \alpha J(r, \pi)$, i.e., for any (r', π'), we have*

$$\begin{cases} \ell(\widehat{r}, \mathcal{D}) + \alpha J(\widehat{r}, \widehat{\pi}) \le \ell(r', \mathcal{D}) + \alpha J(r', \widehat{\pi}) \\ \ell(\widehat{r}, \mathcal{D}) + \alpha J(\widehat{r}, \widehat{\pi}) \ge \ell(\widehat{r}, \mathcal{D}) + \alpha J(\widehat{r}, \pi') \end{cases}.$$

As such, the policy obtained by VPO can be equivalently written as

$$\widehat{\pi} \in \arg\max_\pi \min_r \left\{ J(r, \pi) + \frac{1}{\alpha}\ell(r, \mathcal{D}) \right\} = \arg\max_\pi \min_{r \in \mathcal{R}_{\delta(\pi,\alpha)}} J(r, \pi), \tag{8.24}$$

where $\mathcal{R}_{\delta(\pi,\alpha)}$ is the constraint set $\{r : \ell(r, \mathcal{D}) \le \ell(r_{\mathsf{MLE}}, \mathcal{D}) + \delta(\pi, \alpha)\}$ such that the constrained optimization problem $\min_{r \in \mathcal{R}_{\delta(\pi,\alpha)}} J(r, \pi)$ is equivalent to the regularized problem $\min_r \big\{J(r, \pi) + \frac{1}{\alpha}\ell(r, \mathcal{D})\big\}$. In view of the similarity between the formulations (8.22) and (8.24), we conclude that VPO implements the pessimism principle (8.22) in an oblivious manner without explicitly estimating the uncertainty level, justifying popular practice as a valid approach to pessimism [Kumar et al., 2020].

Theoretical analysis. The next theorem establishes the sub-optimality gap of VPO with linear function approximation under mild assumptions. The proof is given in Appendix G.2.

Theorem 14. *Under Assumptions 7 and 8, let $\widehat{\theta} \in \Theta$ denote the corresponding reward model for $\widehat{\pi}$. Assume that $\|\theta^\star\|_2 \le C$ and $\|\widehat{\theta}\|_2 \le C$ for some $C > 0$. Let $\alpha = \sqrt{N}$ and $\delta \in (0, 1)$. With probability $1 - \delta$, we have*

$$J^\star(r^\star) - J(r^\star, \widehat{\pi}) \le \mathcal{O}\left(\frac{C_1}{\sqrt{N}} \cdot \Big\| \mathop{\mathbb{E}}_{\substack{x \sim \rho, \\ y \sim \pi^\star(\cdot|x)}} [\phi(x, y)] \Big\|_{(\Sigma_\mathcal{D} + \lambda I)^{-1}} + \frac{C_2}{\sqrt{N}} \right),$$

where $\Sigma_\mathcal{D} = \frac{1}{N}\sum_{i=1}^N (\phi(x^i, y_+^i) - \phi(x^i, y_-^i))(\phi(x^i, y_+^i) - \phi(x^i, y_-^i))^\top$ is the feature sample covariance matrix, $\lambda = 1/N$, $C_1 = \exp(C)\Big(\sqrt{d + \log(1/\delta)} + \kappa_\mathcal{D}\Big) + C$ and $C_2 = \exp(C)\kappa_\mathcal{D}^2 + C\kappa_\mathcal{D} + 1$. Here,

$$\kappa_\mathcal{D} = \Big\| \mathop{\mathbb{E}}_{\substack{x \sim \rho, \\ y \sim \widehat{\pi}(\cdot|x)}} [\phi(x, y)] - \mathop{\mathbb{E}}_{\substack{x \sim \rho, \\ y \sim \pi_{cal}(\cdot|x)}} [\phi(x, y)] \Big\|_{(\Sigma_\mathcal{D} + \lambda I)^{-1}} \le 4(\lambda_{\min}(\Sigma_\mathcal{D}) + \lambda)^{-1}.$$

Theorem 14 establishes that VPO achieves the same rate of $\widetilde{\mathcal{O}}(1/\sqrt{N})$ as standard offline RL, as long as the offline dataset $\mathcal{D}$ has sufficient coverage. We remark that $\Big\| \mathop{\mathbb{E}}_{\substack{x \sim \rho, \\ y \sim \pi^\star(\cdot|x)}} [\phi(x, y)] \Big\|_{(\Sigma_\mathcal{D} + \lambda I)^{-1}}$ is reminiscent of the standard single-policy concentratability coefficient in offline RL, which measures the distribution shift between the offline dataset and the optimal policy [Zhu et al., 2023].

8.3 Experiments

In this section, we evaluate the proposed VPO on synthetic multi-armed bandit (MAB) in online and offline settings.

8.3.1 Synthetic multi-armed bandits

We evaluate the proposed methods on a synthetic dataset of size $|\mathcal{X}| = 1$ and $|\mathcal{Y}| = 10$. We set $\pi_{\text{ref}} = \pi_{\mathsf{b}} = \pi_{\text{cal}}$, where $\pi_{\text{ref}} = \mathsf{softmax}(\theta_{\text{ref}})$ with $\theta_{\text{ref}}(x, y)$ sampled i.i.d. from $\mathcal{N}(0, 1)$. The ground truth reward $r^\star$ is randomly generated i.i.d. according to $r^\star(x, y) \sim U([0, 1])$. We approximately solve the optimization problems by performing 20 AdamW optimization steps with learning rate 0.01 and weight decay rate 0.01 in every iteration for the online setting and 1000 steps for the offline setting.

We plot the average results over 10 independent runs in Figure 8.1. As demonstrated in the left panel of Figure 8.1, an appropriate choice of α allows our method to outperform the model-based MAB with MLE baseline in the long-term performance of cumulative regret, at the cost of slightly increased cumulative regret in the first 100 iterations. This highlights the effectiveness of the VPO in achieving more principled exploration-exploitation trade-off. For the offline setting, the right panel of Figure 8.1 demonstrates that the performance of both MLE-MAB and VPO improves as the number of offline data increases. However, VPO achieves a consistently lower sub-optimality gap compared with that of MLE-MAB.

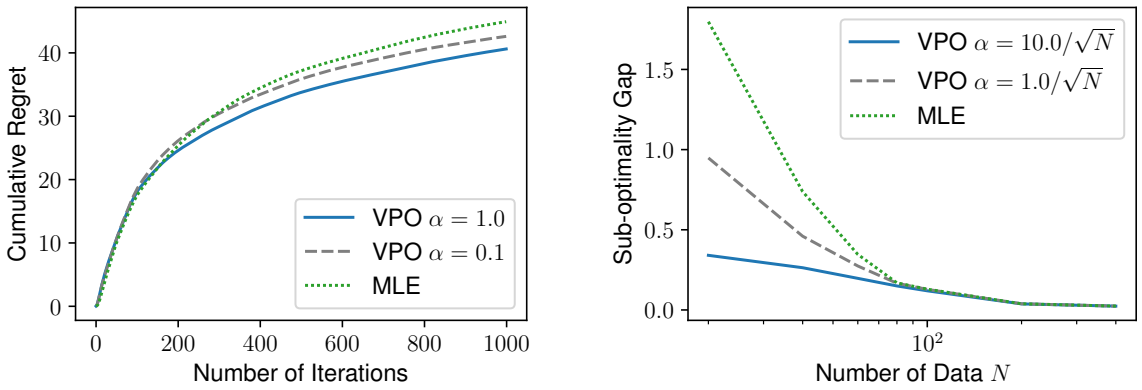


Figure 8.1: The cumulative regret v.s. number of iterations plot (left panel) and sub-optimality gap v.s. number of data plot (right panel) of VPO and MLE-MAB methods in the online and offline settings, respectively.

8.4 Discussion

In this work, we develop a unified approach to achieving principled optimism and pessimism in online and offline RLHF, which enables a practical computation scheme by incorporating uncertainty estimation implicitly within reward-biased maximum likelihood estimation. Theoretical analysis indicates that the proposed methods mirror the guarantees of their standard RL counterparts, which is furthermore corroborated by numerical results. Important future directions include investigating adaptive rules for selecting α without prior information and more refined analysis on the

choice of π_{cal}. This work also hints at a general methodology of designing practical algorithms with principled optimism/pessimism under more general RL setups.

Appendix A

Proofs for Chapter 2

A.1 Analysis

A.1.1 Main pillars for the convergence analysis

Before proceeding, we isolate a few ingredients that provide the main pillars for our theoretical development.

Performance improvement and monotonicity. This lemma is a sort of *ascent lemma*, which quantifies the progress made over each iteration — measured in terms of the soft value function.

Lemma 3 (Performance improvement)**.** *Suppose that* $0 < \eta \leq (1-\gamma)/\tau$*. For any distribution* ρ*, one has*

$$V_\tau^{(t+1)}(\rho) - V_\tau^{(t)}(\rho) = \mathop{\mathbb{E}}_{s\sim d_\rho^{(t+1)}}\left[\left(\frac{1}{\eta} - \frac{\tau}{1-\gamma}\right)\mathsf{KL}\left(\pi^{(t+1)}(\cdot|s)\,\|\,\pi^{(t)}(\cdot|s)\right) + \frac{1}{\eta}\mathsf{KL}\left(\pi^{(t)}(\cdot|s)\,\|\,\pi^{(t+1)}(\cdot|s)\right)\right]. \tag{A.1}$$

Proof. See Appendix A.3.1. □

In a nutshell, Lemma 3 asserts that each iteration of the entropy-regularized NPG method is guaranteed to improve the estimates of the soft value function, with the improvement depending on the KL divergence between the current policy $\pi^{(t)}$ and the updated one $\pi^{(t+1)}$. In fact, the arbitrary choice of ρ readily reveals a sort of *pointwise* monotoncity for the above range of learning rates, in the sense that $V_\tau^{(t+1)}(s) \geq V_\tau^{(t)}(s)$ for all $s \in \mathcal{S}$. Indeed, this lemma can be viewed as the counterpart of the performance difference lemma in Kakade and Langford [2002] for the unregularized form. Lemma 3 also implies the monotonicity of the soft Q-function in t, since for any $(s,a) \in \mathcal{S}\times\mathcal{A}$ one has

$$Q_\tau^{(t+1)}(s,a) = r(s,a) + \gamma \mathop{\mathbb{E}}_{s'\sim P(\cdot|s,a)}\left[V_\tau^{(t+1)}(s')\right] \geq r(s,a) + \gamma \mathop{\mathbb{E}}_{s'\sim P(\cdot|s,a)}\left[V_\tau^{(t)}(s')\right] = Q_\tau^{(t)}(s,a), \tag{A.2}$$

where the equalities follow from the definition (2.10a), and the inequality follows since $V_\tau^{(t+1)}(s) \geq V_\tau^{(t)}(s)$ for all $s \in \mathcal{S}$ — a consequence of Lemma 3 and the non-negativity of the KL divergence.

A key contraction operator: the soft Bellman optimality operator. An operator that plays a pivotal role in the theory of dynamic programming [Bellman, 1952] is the renowned Bellman optimality operator $\mathcal{T} : \mathbb{R}^{|\mathcal{S}||\mathcal{A}|} \to \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$, defined as follows

$$\forall (s,a) \in \mathcal{S} \times \mathcal{A} : \quad \mathcal{T}(Q)(s,a) := r(s,a) + \gamma \mathop{\mathbb{E}}_{s' \sim P(\cdot|s,a)} \left[\max_{a'} Q(s',a') \right]. \tag{A.3}$$

In order to facilitate analysis for entropy-regularized MDPs, we find it particularly fruitful to introduce a "soft" Bellman optimality operator $\mathcal{T}_\tau : \mathbb{R}^{|\mathcal{S}||\mathcal{A}|} \to \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ as follows

$$\forall (s,a) \in \mathcal{S} \times \mathcal{A} : \quad \mathcal{T}_\tau(Q)(s,a) := r(s,a) + \gamma \mathop{\mathbb{E}}_{s' \sim P(\cdot|s,a)} \left[\max_{\pi(\cdot|s') \in \Delta(\mathcal{A})} \mathop{\mathbb{E}}_{a' \sim \pi(\cdot|s')} \left[Q(s',a') - \tau \log \pi(a'|s') \right] \right], \tag{A.4}$$

which reduces to $\mathcal{T}$ when $\tau = 0$. To see this, observe that

$$\begin{aligned} \mathcal{T}_0(Q)(s,a) &= r(s,a) + \gamma \mathop{\mathbb{E}}_{s' \sim P(\cdot|s,a)} \left[\max_{\pi(\cdot|s') \in \Delta(\mathcal{A})} \mathop{\mathbb{E}}_{a' \sim \pi(\cdot|s')} \left[Q(s',a') \right] \right] \\ &= r(s,a) + \gamma \mathop{\mathbb{E}}_{s' \sim P(\cdot|s,a)} \left[\max_{a'} Q(s',a') \right] = \mathcal{T}(Q)(s,a), \end{aligned}$$

where the last line follows since the optimal policy is exactly the greedy policy w.r.t. Q [Puterman, 2014]. The operator $\mathcal{T}_\tau$ plays a similar role as does the Bellman optimality operator for the unregularized case, whose key properties are summarized below. Similar results have been derived in Dai et al. [2018, Section 3.1].

Lemma 4 (Soft Bellman optimality operator)**.** *The operator $\mathcal{T}_\tau$ defined in (A.4) satisfies the properties below.*

- *$\mathcal{T}_\tau$ admits the following closed-form expression:*

$$\mathcal{T}_\tau(Q)(s,a) = r(s,a) + \gamma \mathop{\mathbb{E}}_{s' \sim P(\cdot|s,a)} \left[\tau \log \left(\left\| \exp\left(Q^{(()} s', \cdot)/\tau \right) \right\|_1 \right) \right]. \tag{A.5}$$

- *The optimal soft Q-function $Q_\tau^\star$ is a fixed point of $\mathcal{T}_\tau$, namely,*

$$\mathcal{T}_\tau\big(Q_\tau^\star\big) = Q_\tau^\star. \tag{A.6}$$

- *$\mathcal{T}_\tau$ is a γ-contraction in the ℓ_∞ norm, namely, for any $Q_1, Q_2 \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ one has*

$$\left\| \mathcal{T}_\tau(Q_1) - \mathcal{T}_\tau(Q_2) \right\|_\infty \le \gamma \left\| Q_1 - Q_2 \right\|_\infty. \tag{A.7}$$

Proof. See Appendix A.3.2. □

For those familiar with dynamic programming, it should become evident that $\mathcal{T}_\tau$ inherits many appealing features of the original Bellman optimality operator $\mathcal{T}$. For example, as an immediate application of the γ-contraction property (A.7) and the fixed-point property (A.6), the following soft Q-value iteration

$$Q_{\mathsf{svi}}^{(t+1)} \;=\; \mathcal{T}_\tau\big(Q_{\mathsf{svi}}^{(t)}\big), \qquad t \ge 0$$

is guaranteed to converge linearly to the optimal $Q_\tau^\star$ with a contraction rate γ — a simple observation consistent with the behavior of value iteration designed for unregularized MDPs.

A.1.2 Analysis of exact entropy-regularized NPG methods

The SPI case (i.e. $\eta = (1-\gamma)/\tau$)

With the help of the soft Bellman optimality operator, we have

$$\begin{aligned} Q_\tau^{(t+1)}(s,a) &\overset{\text{(i)}}{=} r(s,a) + \gamma \mathop{\mathbb{E}}_{s'\sim P(\cdot|s,a)} \Big[V_\tau^{(t+1)}(s')\Big] \\ &\overset{\text{(ii)}}{=} r(s,a) + \gamma \mathop{\mathbb{E}}_{\substack{s'\sim P(\cdot|s,a),\\ a'\sim\pi^{(t+1)}(\cdot|s')}} \Big[-\tau\log\pi^{(t+1)}(a'|s') + Q_\tau^{(t+1)}(s',a')\Big] \\ &\overset{\text{(iii)}}{\geq} r(s,a) + \gamma \mathop{\mathbb{E}}_{\substack{s'\sim P(\cdot|s,a),\\ a'\sim\pi^{(t+1)}(\cdot|s')}} \Big[-\tau\log\pi^{(t+1)}(a'|s') + Q_\tau^{(t)}(s',a')\Big] \\ &\overset{\text{(iv)}}{=} r(s,a) + \gamma \mathop{\mathbb{E}}_{s'\sim P(\cdot|s,a)} \Big[\tau\log\Big(\big\|\exp\big(Q^{(t)}(s',\cdot)/\tau\big)\big\|_1\Big)\Big] \\ &\overset{\text{(v)}}{=} \mathcal{T}_\tau\big(Q_\tau^{(t)}\big)(s,a). \end{aligned} \tag{A.8}$$

Here, (i) comes from the definition (2.10a) of the soft Q-function, (ii) follows from the relation (2.10b), (iii) relies on the monotonicity of the soft Q-function (see (A.2)), (iv) uses the form of $\pi^{(t+1)}$ in (2.16), whereas (v) makes use of the expression (A.5). The inequality (A.8) further leads to $0 \leq Q_\tau^\star - Q_\tau^{(t+1)} \leq Q_\tau^\star - \mathcal{T}_\tau\big(Q_\tau^{(t+1)}\big)$, and hence

$$\begin{aligned} \big\|Q_\tau^\star - Q_\tau^{(t+1)}\big\|_\infty &\leq \big\|Q_\tau^\star - \mathcal{T}_\tau\big(Q_\tau^{(t)}\big)\big\|_\infty = \big\|\mathcal{T}_\tau\big(Q_\tau^\star\big) - \mathcal{T}_\tau\big(Q_\tau^{(t)}\big)\big\|_\infty \leq \gamma\big\|Q_\tau^\star - Q_\tau^{(t)}\big\|_\infty \\ &\leq \gamma^{t+1}\big\|Q_\tau^\star - Q_\tau^{(0)}\big\|_\infty, \end{aligned} \tag{A.9}$$

where the first equality follows from the fixed-point property (A.6), and the second inequality is due to the contraction property (A.7). We have thus established linear convergence of $Q_\tau^{(t)}$ in $\|\cdot\|_\infty$ for this case.

Turning to the log policies, recall that

$$\pi^{(t+1)}(\cdot|s) \propto \exp\big(Q_\tau^{(t)}(s,\cdot)/\tau\big) \qquad \text{and} \qquad \pi_\tau^\star(\cdot|s) \propto \exp\big(Q_\tau^\star(s,\cdot)/\tau\big),$$

where the second relation comes from Nachum et al. [2017, Eqn. (12)]. It then follows from an elementary property of the softmax function (see (A.34) in Appendix A.2.2) that

$$\big\|\log\pi^{(t+1)} - \log\pi_\tau^\star\big\|_\infty \leq \frac{2}{\tau}\big\|Q_\tau^{(t)} - Q_\tau^\star\big\|_\infty \leq \frac{2}{\tau}\gamma^t\big\|Q_\tau^\star - Q_\tau^{(0)}\big\|_\infty,$$

thus concluding the proof for this case.

The case with general learning rates

We now move to the case with a general learning rate. For the sake of brevity, we shall denote

$$\alpha := 1 - \frac{\eta\tau}{1-\gamma}. \tag{A.10}$$

Additionally, it is helpful to introduce an auxiliary sequence $\{\xi^{(t)} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}\}$ constructed recursively by

$$\xi^{(0)}(s,a) := \big\|\exp\big(Q_\tau^\star(s,\cdot)/\tau\big)\big\|_1 \cdot \pi^{(0)}(a|s), \tag{A.11a}$$

$$\xi^{(t+1)}(s,a) := \big[\xi^{(t)}(s,a)\big]^\alpha \exp\Big((1-\alpha)\frac{Q_\tau^{(t)}(s,a)}{\tau}\Big), \qquad \forall\ (s,a)\in\mathcal{S}\times\mathcal{A},\ t\geq 0. \tag{A.11b}$$

It is easily seen from the construction (A.11b) that

$$\begin{aligned} Q_\tau^\star - \tau \log \xi^{(t+1)} &= Q_\tau^\star - \tau\alpha \log \xi^{(t)} - (1-\alpha) Q_\tau^{(t)} \\ &= \alpha\big(Q_\tau^\star - \tau \log \xi^{(t)}\big) + (1-\alpha)\big(Q_\tau^\star - Q_\tau^{(t)}\big) \end{aligned} \tag{A.12}$$

and, consequently,

$$\big\|Q_\tau^\star - \tau \log \xi^{(t+1)}\big\|_\infty \le \alpha \big\|Q_\tau^\star - \tau \log \xi^{(t)}\big\|_\infty + (1-\alpha)\big\|Q_\tau^\star - Q_\tau^{(t)}\big\|_\infty. \tag{A.13}$$

Step 1: a linear system that describes the error recursions. In the case with general learning rates, the estimation error $\big\|Q_\tau^\star - Q_\tau^{(t)}\big\|_\infty$ does not contract in the same form as that of soft policy iteration; instead, it is more succinctly controlled with the aid of an auxiliary quantity $\big\|Q_\tau^\star - \tau \log \xi^{(t)}\big\|_\infty$. In what follows, we leverage a simple yet powerful technique by describing the dynamics concerning $\big\|Q_\tau^\star - Q_\tau^{(t)}\big\|_\infty$ and $\big\|Q_\tau^\star - \tau \log \xi^{(t)}\big\|_\infty$ via a linear system, whose spectral properties dictate the convergence rate. Towards this, we start with the following key observation, whose proof is deferred to Appendix A.3.3.

Lemma 5. *For any learning rate $0 < \eta \le (1-\gamma)/\tau$, the entropy-regularized NPG updates (2.17) satisfy*

$$\big\|Q_\tau^\star - Q_\tau^{(t+1)}\big\|_\infty \le \gamma \big\|Q_\tau^\star - \tau \log \xi^{(t+1)}\big\|_\infty + \gamma\alpha^{t+1}\big\|Q_\tau^{(0)} - \tau \log \xi^{(0)}\big\|_\infty, \tag{A.14}$$

where α is defined in (A.10).

If we substitute (A.12) into (A.14), it is straightforwardly seen that Lemma 5 is a generalization of the contraction property (A.9) of soft policy iteration (the case corresponding to $\alpha = 0$). Given that Lemma 5 involves the interaction of more than one quantities, it is convenient to combine (A.13) and (A.14) into the following linear system

$$x_{t+1} \le A x_t + \gamma \alpha^{t+1} y, \tag{A.15}$$

where

$$A := \begin{bmatrix} \gamma(1-\alpha) & \gamma\alpha \\ 1-\alpha & \alpha \end{bmatrix}, \quad x_t := \begin{bmatrix} \big\|Q_\tau^\star - Q_\tau^{(t)}\big\|_\infty \\ \big\|Q_\tau^\star - \tau \log \xi^{(t)}\big\|_\infty \end{bmatrix} \quad \text{and} \quad y := \begin{bmatrix} \big\|Q_\tau^{(0)} - \tau \log \xi^{(0)}\big\|_\infty \\ 0 \end{bmatrix}. \tag{A.16}$$

We shall make note of the following appealing features of the rank-1 system matrix A:

$$A = \begin{bmatrix} \gamma \\ 1 \end{bmatrix} \big[1-\alpha,\, \alpha\big], \qquad \text{and} \quad A^t = (1-\eta\tau)^{t-1} A \qquad \forall t \ge 0, \tag{A.17}$$

which relies on the identity $(1-\alpha)\gamma + \alpha = 1 - \eta\tau$ (according to the definition (A.10) of α).

Remark 11. *By left multiplying both sides of (A.15) by $[1-\alpha, \alpha]$, we obtain*

$$L^{(t+1)} \le (1-\eta\tau) L^{(t)} + \gamma(1-\alpha)\alpha^{t+1}\big\|Q_\tau^{(0)} - \tau \log \xi^{(0)}\big\|_\infty,$$

where $L^{(t)} := (1-\alpha)\big\|Q_\tau^\star - Q_\tau^{(t)}\big\|_\infty + \alpha\big\|Q_\tau^\star - \tau \log \xi^{(t)}\big\|_\infty$ can be viewed as a sort of Lyapunov function. This hints at the intimate connection between our proof and the Lyapunov-type analysis used in system theory.

Step 2: characterizing the contraction rate from the linear system. In view of the recursion formula (A.15) and the non-negativity of (A, x_t, y), it is immediate to deduce that

$$\begin{aligned} x_{t+1} &\le A(Ax_{t-1} + \gamma\alpha^t y) + \gamma\alpha^{t+1} y \\ &\le A^{t+1}x_0 + \gamma\left(\alpha^{t+1}I + \alpha^t A + \cdots + \alpha A^t\right) y \\ &= A^{t+1}x_0 + \gamma\left(A^{t+1} - \alpha^{t+1}I\right)\left(\alpha^{-1}A - I\right)^{-1} y. \end{aligned} \tag{A.18}$$

Here, the last line follows from the elementary relation

$$\left(\alpha^{t+1}I + \alpha^t A + \cdots + \alpha A^t\right)\left(\alpha^{-1}A - I\right) = A^{t+1} - \alpha^{t+1}I$$

and the invertibility of $\alpha^{-1}A - I$ (since $\alpha^{-1}A$ is a rank-1 matrix whose non-zero singular value is larger than 1). In addition, the Woodbury matrix inversion formula together with the decomposition (A.17) yields

$$\gamma\left(\alpha^{-1}A - I\right)^{-1} y = \gamma\left\{\begin{bmatrix} 1 & \frac{\alpha}{1-\alpha} \\ \frac{1}{\gamma} & \frac{\alpha}{(1-\alpha)\gamma} \end{bmatrix} - I\right\} y = \begin{bmatrix} 0 & \frac{\gamma\alpha}{1-\alpha} \\ 1 & \frac{\gamma\alpha+\alpha-\gamma}{1-\alpha} \end{bmatrix} y = \begin{bmatrix} 0 \\ \left\|Q_\tau^{(0)} - \tau\log\xi^{(0)}\right\|_\infty \end{bmatrix}, \tag{A.19}$$

which is a non-negative vector. Consequently, this taken together with (A.18) gives

$$\begin{aligned} x_{t+1} &\le A^{t+1}\left[x_0 + \gamma\left(\alpha^{-1}A - I\right)^{-1} y\right] - \alpha^{t+1}\left\{\gamma\left(\alpha^{-1}A - I\right)^{-1} y\right\} \\ &\le A^{t+1}\left[x_0 + \gamma\left(\alpha^{-1}A - I\right)^{-1} y\right] \\ &= (1-\eta\tau)^t\left(\begin{bmatrix}\gamma \\ 1\end{bmatrix}\begin{bmatrix}1-\alpha, \alpha\end{bmatrix}\right)\begin{bmatrix} \left\|Q_\tau^\star - Q_\tau^{(0)}\right\|_\infty \\ \left\|Q_\tau^\star - \tau\log\xi^{(0)}\right\|_\infty + \left\|Q_\tau^{(0)} - \tau\log\xi^{(0)}\right\|_\infty \end{bmatrix} \\ &= (1-\eta\tau)^t\left\{(1-\alpha)\left\|Q_\tau^\star - Q_\tau^{(0)}\right\|_\infty + \alpha\left(\left\|Q_\tau^\star - \tau\log\xi^{(0)}\right\|_\infty + \left\|Q_\tau^{(0)} - \tau\log\xi^{(0)}\right\|_\infty\right)\right\}\begin{bmatrix}\gamma \\ 1\end{bmatrix}, \end{aligned} \tag{A.20}$$

where the third line follows from (A.17), (A.19) and the definition of x_t. Further, observe that

$$\begin{aligned} &\left\|Q_\tau^\star - \tau\log\xi^{(0)}\right\|_\infty + \left\|Q_\tau^{(0)} - \tau\log\xi^{(0)}\right\|_\infty - \left\|Q_\tau^\star - Q_\tau^{(0)}\right\|_\infty \\ &\qquad \le 2\left\|Q_\tau^\star - \tau\log\xi^{(0)}\right\|_\infty = 2\tau\left\|\log\pi_\tau^\star - \log\pi^{(0)}\right\|_\infty, \end{aligned} \tag{A.21}$$

where the inequality comes from the triangle inequality, and the last identity follows from (A.11a). Substituting this back into (A.20), we obtain

$$x_{t+1} \le (1-\eta\tau)^t\left\{\left\|Q_\tau^\star - Q_\tau^{(0)}\right\|_\infty + 2\alpha\tau\left\|\log\pi_\tau^\star - \log\pi^{(0)}\right\|_\infty\right\}\begin{bmatrix}\gamma \\ 1\end{bmatrix}. \tag{A.22}$$

To finish up, recall that $\pi^{(t)}$ is related to $\xi^{(t)}$ as follows

$$\forall s \in \mathcal{S}: \qquad \pi^{(t)}(\cdot|s) = \frac{1}{\|\xi^{(t)}(s,\cdot)\|_1}\xi^{(t)}(s,\cdot), \tag{A.23}$$

which can be seen by comparing (A.11) with (2.17). Therefore, invoking the elementary property of the softmax function (see (A.34) in Appendix A.2.2), we arrive at

$$\left\|\log\pi_\tau^\star - \log\pi^{(t+1)}\right\|_\infty \le 2\left\|Q_\tau^\star/\tau - \log\xi^{(t+1)}\right\|_\infty.$$

This combined with (A.22) as well as the definition (A.16) of x_{t+1} immediately establishes Theorem 1.

A.1.3 Analysis of approximate entropy-regularized NPG methods

We now turn to the convergence properties of approximate entropy-regularized NPG methods — as claimed in Theorem 2 — when only inexact policy evaluation $\widehat{Q}_\tau^{(t)}$ is available (in the sense of (2.22)).

Step 1: performance difference accounting for inexact policy evaluation. We first bound the quality of the policy updates (2.22) by examining the difference between $V_\tau^{(t+1)}$ and $V_\tau^{(t)}$ and how it is impacted by the imperfectness of policy evaluation. This is made precise by the following lemma.

Lemma 6 (Performance difference of approximate entropy-regularized NPG)**.** *Suppose that* $0 < \eta \le (1-\gamma)/\tau$. *For any state* $s_0 \in \mathcal{S}$, *one has*

$$V_\tau^{(t)}(s_0) \le V_\tau^{(t+1)}(s_0) + \frac{2}{1-\gamma}\big\|\widehat{Q}_\tau^{(t)} - Q_\tau^{(t)}\big\|_\infty. \tag{A.24}$$

Proof. See Appendix A.3.4. □

The careful reader might already realize that the above lemma is a relaxation of Lemma 3; in particular, the last term of (A.24) quantifies the effect of the approximation error (i.e. the difference between $\widehat{Q}_\tau^{(t)}$ and $Q_\tau^{(t)}$) upon performance improvement. Under the assumption $\big\|\widehat{Q}_\tau^{(t)} - Q_\tau^{(t)}\big\|_\infty \le \delta$, repeating the argument of (A.2) reveals that the soft Q-function estimates are not far from being monotone in t, in the sense that

$$\forall (s,a) \in \mathcal{S}\times\mathcal{A}: \qquad Q_\tau^{(t)}(s,a) - Q_\tau^{(t+1)}(s,a) = \gamma \mathop{\mathbb{E}}_{s'\sim P(\cdot|s,a)}\Big[V_\tau^{(t)}(s') - V_\tau^{(t+1)}(s')\Big] \le \frac{2\gamma\delta}{1-\gamma}. \tag{A.25}$$

Step 2: a linear system accounting for inexact policy evaluation. With the assistance of (A.25), it is possible to construct a linear system — similar to the one built in Section A.1.2 — that takes into account inexact policy evaluation. Towards this end, we adopt a similar approach as in (A.11) by introducing the following auxiliary sequence $\widehat{\xi}^{(t)}$ defined recursively using $\widehat{Q}_\tau^{(t)}$:

$$\widehat{\xi}^{(0)}(s,a) := \big\|\exp\big(Q_\tau^\star(s,\cdot)/\tau\big)\big\|_1 \cdot \pi^{(0)}(s,a), \tag{A.26a}$$

$$\widehat{\xi}^{(t+1)}(s,a) := \Big[\widehat{\xi}^{(t)}(s,a)\Big]^\alpha \exp\Big((1-\alpha)\frac{\widehat{Q}_\tau^{(t)}(s,a)}{\tau}\Big), \qquad \forall\ (s,a)\in\mathcal{S}\times\mathcal{A},\ t\ge 0, \tag{A.26b}$$

where $\alpha := 1 - \frac{\eta\tau}{1-\gamma}$ as before.

We claim that the following linear system tracks the error dynamics of the policy updates:

$$z_{t+1} \le B z_t + b, \tag{A.27}$$

where

$$B := \begin{bmatrix} \gamma(1-\alpha) & \gamma\alpha & \gamma\alpha \\ 1-\alpha & \alpha & 0 \\ 0 & 0 & \alpha \end{bmatrix},\ z_t := \begin{bmatrix} \big\|Q_\tau^\star - Q_\tau^{(t)}\big\|_\infty \\ \big\|Q_\tau^\star - \tau\log\widehat{\xi}^{(t)}\big\|_\infty \\ -\min_{s,a}\big(Q_\tau^{(t)}(s,a) - \tau\log\widehat{\xi}^{(t)}(s,a)\big) \end{bmatrix},$$

$$b := (1-\alpha)\delta\begin{bmatrix} \gamma\left(2+\frac{2\gamma}{\eta\tau}\right) \\ 1 \\ 1+\frac{2\gamma}{\eta\tau} \end{bmatrix}. \tag{A.28}$$

Here, the system matrix B (in particular its eigenvalues) governs the contraction rate, while the term b captures the error introduced by inexact policy evaluation. Theorem 2 then follows by carrying out a similar analysis argument as in Section A.1.2 to characterize the error dynamics. Details are postponed to Appendix A.4.

A.2 Preliminaries

A.2.1 Derivation of entropy-regularized NPG methods

This subsection establishes the equivalence between the update rules (2.14) and (2.17). Such derivations are inherently similar to the ones for the NPG update rule (without entropy regularization) (see, e.g., Agarwal et al. [2019]); we provide the proof here for pedagogical reasons.

First of all, let us follow the convention to introduce the advantage function $A_\tau^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ of a policy π w.r.t. the entropy-regularized value function:

$$\forall (s,a) \in \mathcal{S} \times \mathcal{A}: \qquad A_\tau^\pi(s,a) := Q_\tau^\pi(s,a) - \tau \log \pi(a|s) - V_\tau^\pi(s) \tag{A.29}$$

with Q_τ^π defined in (2.10a), which reflects the gain one can harvest by executing action a instead of following the policy π in state s. This advantage function plays a crucial role in the calculation of policy gradients, due to the following fundamental relation (see Appendix A.3.5 for the proof):

Lemma 7. *Under softmax parameterization (2.6), the gradient of the regularized value function satisfies*

$$\frac{\partial V_\tau^{\pi_\theta}(\rho)}{\partial \theta(s,a)} = \frac{1}{1-\gamma} d_\rho^{\pi_\theta}(s) \cdot \pi_\theta(a|s) \cdot A_\tau^{\pi_\theta}(s,a); \tag{A.30a}$$

$$\left[\left(\mathcal{F}_\rho^\theta\right)^\dagger \nabla_\theta V_\tau^{\pi_\theta}(\rho) \right](s,a) = \frac{1}{1-\gamma} A_\tau^{\pi_\theta}(s,a) + c(s) \tag{A.30b}$$

for any $(s,a) \in \mathcal{S} \times \mathcal{A}$, where $c(s) := \sum_a \pi_\theta(a|s) w_{s,a}$ is some function depending only on s.

It is worth highlighting that the search direction of NPG, given in (A.30b), is invariant to the choice of ρ. With the above calculations in place, it is seen that for any $s \in \mathcal{S}$, the regularized NPG update rule (2.14) results in a policy update as follows

$$\begin{aligned}
\pi^{(t+1)}(a|s) &\overset{\text{(i)}}{\propto} \exp\left(\theta^{(t+1)}(s,a)\right) \overset{\text{(ii)}}{=} \exp\left(\theta^{(t)}(s,a) + \eta \left[\left(\mathcal{F}_\rho^{\theta^{(t)}}\right)^\dagger \nabla_\theta V_\tau^{(t)}(\rho) \right](s,a)\right) \\
&\overset{\text{(iii)}}{\propto} \exp\left(\theta^{(t)}(s,a) + \frac{\eta}{1-\gamma} A_\tau^{(t)}(s,a)\right) \\
&\overset{\text{(iv)}}{\propto} \pi^{(t)}(a|s) \exp\left(\frac{\eta}{1-\gamma} Q_\tau^{(t)}(s,a) - \frac{\eta\tau}{1-\gamma} \log \pi^{(t)}(a|s)\right) \\
&= \left(\pi^{(t)}(a|s)\right)^{1-\frac{\eta\tau}{1-\gamma}} \exp\left(\frac{\eta}{1-\gamma} Q_\tau^{(t)}(s,a)\right).
\end{aligned}$$

where we use $A_\tau^{(t)}$ to abbreviate $A_\tau^{\pi^{(t)}}$. Here, (i) uses the definition of the softmax policy, (ii) comes from the update rule (2.14), (iii) is a consequence of (A.30b) (since $c(\cdot)$ does not depend on a), whereas (iv) results from the definition (A.29) and the fact that $V_\tau^\pi(\cdot)$ is not dependent on a. This validates the equivalence between (2.14) and (2.17).

A.2.2 Basic facts about the function $\log(\|\exp(\theta)\|_1)$

In the current paper, we often encounter the function $\log\left(\|\exp(\theta)\|_1\right) := \log\left(\sum_{1\le a\le|\mathcal{A}|}\exp(\theta_a)\right)$ for any vector $\theta = [\theta_a]_{1\le a\le|\mathcal{A}|} \in \mathbb{R}^{|\mathcal{A}|}$. To facilitate analysis, we single out several basic properties concerning this function, which will be used multiple times when establishing our main results. For notational convenience, we denote by $\pi_\theta \in \mathbb{R}^{|\mathcal{A}|}$ the softmax transform of θ such that

$$\pi_\theta(a) = \frac{\exp(\theta_a)}{\sum_{1\le j\le|\mathcal{A}|}\exp(\theta_j)}, \qquad 1 \le a \le |\mathcal{A}|. \tag{A.31}$$

By straightforward calculations, the gradient of the function $\log\left(\|\exp(\theta)\|_1\right)$ is given by

$$\nabla_\theta \log\left(\|\exp(\theta)\|_1\right) = \frac{1}{\|\exp(\theta)\|_1}\exp(\theta) = \pi_\theta;. \tag{A.32}$$

Difference of log policies. In the analysis, we often need to control the difference of two policies, towards which the following bounds prove useful. To begin with, the mean value theorem reveals a Lipschitz continuity property (w.r.t. the ℓ_∞ norm): for any $\theta_1, \theta_2 \in \mathbb{R}^{|\mathcal{A}|}$,

$$\begin{aligned}
&\left|\log\left(\|\exp(\theta_1)\|_1\right) - \log\left(\|\exp(\theta_2)\|_1\right)\right| \\
&= \left|\left\langle \theta_1-\theta_2, \nabla_\theta \log\left(\|\exp(\theta)\|_1\right)|_{\theta=\theta_c}\right\rangle\right| \\
&\le \|\theta_1-\theta_2\|_\infty \left\|\nabla_\theta \log\left(\|\exp(\theta)\|_1\right)|_{\theta=\theta_c}\right\|_1 = \|\theta_1-\theta_2\|_\infty,
\end{aligned} \tag{A.33}$$

where θ_c is a certain convex combination of θ_1 and θ_2, and the second line relies on (A.32). In addition, for any two vectors π_{θ_1} and π_{θ_2} defined w.r.t. $\theta_1, \theta_2 \in \mathbb{R}^{|\mathcal{A}|}$ (see (A.31)), one has

$$\|\log\pi_{\theta_1} - \log\pi_{\theta_2}\|_\infty \le 2\|\theta_1-\theta_2\|_\infty, \tag{A.34}$$

where $\log(\cdot)$ denotes entrywise operation. To justify (A.34), we observe from the definition (A.31) that

$$\|\log\pi_{\theta_1} - \log\pi_{\theta_2}\|_\infty \le \|\theta_1-\theta_2\|_\infty + \left|\log\left(\|\exp(\theta_1)\|_1\right) - \log\left(\|\exp(\theta_2)\|_1\right)\right| \le 2\|\theta_1-\theta_2\|_\infty,$$

where the last inequality is a consequence of (A.33).

A.3 Proof for key lemmas

A.3.1 Proof of Lemma 3

To begin with, the regularized NPG update rule (see (2.17) in Algorithm 1) indicates that

$$\log\pi^{(t+1)}(a|s) = \left(1-\frac{\eta\tau}{1-\gamma}\right)\log\pi^{(t)}(a|s) + \frac{\eta}{1-\gamma}Q_\tau^{(t)}(s,a) - \log Z^{(t)}(s), \tag{A.35}$$

where $Z^{(t)}$ is some quantity depending only on the state s (but not the action a). Rearranging terms gives

$$-\tau\log\pi^{(t)}(a|s) + Q_\tau^{(t)}(s,a) = \frac{1-\gamma}{\eta}\left(\log\pi^{(t+1)}(a|s) - \log\pi^{(t)}(a|s)\right) + \frac{1-\gamma}{\eta}\log Z^{(t)}(s). \tag{A.36}$$

This in turn allows us to express $V_\tau^{(t)}(s_0)$ for any $s_0 \in \mathcal{S}$ as follows

$$\begin{aligned}
V_\tau^{(t)}(s_0) &= \mathop{\mathbb{E}}_{a_0\sim\pi^{(t)}(\cdot|s_0)} \left[-\tau \log \pi^{(t)}(a_0|s_0) + Q_\tau^{(t)}(s_0,a_0)\right] \\
&= \mathop{\mathbb{E}}_{a_0\sim\pi^{(t)}(\cdot|s_0)} \left[\frac{1-\gamma}{\eta} \log Z^{(t)}(s_0)\right] + \mathop{\mathbb{E}}_{a_0\sim\pi^{(t)}(\cdot|s_0)} \left[\frac{1-\gamma}{\eta}\left(\log \pi^{(t+1)}(a_0|s_0) - \log \pi^{(t)}(a_0|s_0)\right)\right] \\
&= \frac{1-\gamma}{\eta} \log Z^{(t)}(s_0) - \frac{1-\gamma}{\eta} \mathsf{KL}\left(\pi^{(t)}(\cdot|s_0) \,\|\, \pi^{(t+1)}(\cdot|s_0)\right) \\
&= \mathop{\mathbb{E}}_{a_0\sim\pi^{(t+1)}(\cdot|s_0)} \left[\frac{1-\gamma}{\eta} \log Z^{(t)}(s_0)\right] - \frac{1-\gamma}{\eta} \mathsf{KL}\left(\pi^{(t)}(\cdot|s_0) \,\|\, \pi^{(t+1)}(\cdot|s_0)\right), \qquad \text{(A.37)}
\end{aligned}$$

where the first identity makes use of the definitions (2.7) and (2.10a), the second line follows from (A.36), the third line relies on the definition of the KL divergence, and the last line follows since $Z^{(t)}(s)$ does not depend on a. Invoking (A.36) again to rewrite $\log Z^{(t)}(s_0)$ appearing in the first term of (A.37), we reach

$$\begin{aligned}
&V_\tau^{(t)}(s_0) \\
&= \mathop{\mathbb{E}}_{a_0\sim\pi^{(t+1)}(\cdot|s_0)} \left[-\tau \log \pi^{(t+1)}(a_0|s_0) + Q_\tau^{(t)}(s_0,a_0) + \left(\tau - \frac{1-\gamma}{\eta}\right)\left(\log \pi^{(t+1)}(a_0|s_0) - \log \pi^{(t)}(a_0|s_0)\right)\right] \\
&\quad - \frac{1-\gamma}{\eta} \mathsf{KL}\left(\pi^{(t)}(\cdot|s_0) \,\|\, \pi^{(t+1)}(\cdot|s_0)\right) \\
&= \mathop{\mathbb{E}}_{a_0\sim\pi^{(t+1)}(\cdot|s_0)} \left[-\tau \log \pi^{(t+1)}(a_0|s_0) + Q_\tau^{(t)}(s_0,a_0)\right] + \left(\tau - \frac{1-\gamma}{\eta}\right) \mathsf{KL}\left(\pi^{(t+1)}(\cdot|s_0) \,\|\, \pi^{(t)}(\cdot|s_0)\right) \\
&\quad - \frac{1-\gamma}{\eta} \mathsf{KL}\left(\pi^{(t)}(\cdot|s_0) \,\|\, \pi^{(t+1)}(\cdot|s_0)\right) \\
&= \mathop{\mathbb{E}}_{\substack{a_0\sim\pi^{(t+1)}(\cdot|s_0),\\ s_1\sim P(\cdot|s_0,a_0)}} \left[-\tau \log \pi^{(t+1)}(a_0|s_0) + r(s_0,a_0) + \gamma V_\tau^{(t)}(s_1)\right] \\
&\quad - \left(\frac{1-\gamma}{\eta} - \tau\right) \mathsf{KL}\left(\pi^{(t+1)}(\cdot|s_0) \,\|\, \pi^{(t)}(\cdot|s_0)\right) - \frac{1-\gamma}{\eta} \mathsf{KL}\left(\pi^{(t)}(\cdot|s_0) \,\|\, \pi^{(t+1)}(\cdot|s_0)\right), \qquad \text{(A.38)}
\end{aligned}$$

where the second line uses the definition of the KL divergence, and the third line expands $Q_\tau^{(t)}$ using the definition (2.10a).

To finish up, applying the above relation (A.38) recursively to expand $V_\tau^{(t)}(s_i)$ $(i \geq 1)$, we arrive at

$$\begin{aligned}
V_\tau^{(t)}(s_0) &= \mathop{\mathbb{E}}_{\substack{a_i\sim\pi^{(t+1)}(\cdot|s_i),\\ s_{i+1}\sim P(\cdot|s_i,a_i),\forall i\geq 0}} \left[\sum_{i=0}^{\infty} \gamma^i \left\{r(s_i,a_i) - \tau \log \pi^{(t+1)}(a_i|s_i)\right\}\right. \\
&\qquad \left. - \sum_{i=0}^{\infty} \gamma^i \left\{\left(\frac{1-\gamma}{\eta} - \tau\right) \mathsf{KL}\left(\pi^{(t+1)}(\cdot|s_i) \,\|\, \pi^{(t)}(\cdot|s_i)\right) + \frac{1-\gamma}{\eta} \mathsf{KL}\left(\pi^{(t)}(\cdot|s_i) \,\|\, \pi^{(t+1)}(\cdot|s_i)\right)\right\}\right] \\
&= V_\tau^{(t+1)}(s_0) - \mathop{\mathbb{E}}_{s\sim d_{s_0}^{(t+1)}} \left[\left(\frac{1}{\eta} - \frac{\tau}{1-\gamma}\right) \mathsf{KL}\left(\pi^{(t+1)}(\cdot|s) \,\|\, \pi^{(t)}(\cdot|s)\right) + \frac{1}{\eta} \mathsf{KL}\left(\pi^{(t)}(\cdot|s) \,\|\, \pi^{(t+1)}(\cdot|s)\right)\right], \\
&\qquad \text{(A.39)}
\end{aligned}$$

where the second line follows since the regularized value function $V_\tau^{(t+1)}$ can be viewed as the value function of $\pi^{(t+1)}$ with adjusted rewards $r_\tau^{(t+1)}(s,a) := r(s,a) - \tau \log \pi^{(t+1)}(a|s)$. Averaging the initial state s_0 over the distribution ρ concludes the proof.

A.3.2 Proof of Lemma 4

In the sequel, we prove each claim in Lemma 4 in order.

Proof of Eqn. (A.5). Jensen's inequality tells us that: for any $s \in \mathcal{S}$ one has

$$\begin{aligned}
\mathop{\mathbb{E}}_{a\sim\pi(\cdot|s)}\Big[Q(s,a) - \tau\log\pi(a|s)\Big] &= \tau\sum_a \pi(a|s)\log\left(\frac{\exp\big(Q(s,a)/\tau\big)}{\pi(a|s)}\right) \\
&\le \tau\log\left(\sum_a \pi(a|s)\frac{\exp\big(Q(s,a)/\tau\big)}{\pi(a|s)}\right) \\
&= \tau\log\left(\sum_a \exp\big(Q(s,a)/\tau\big)\right) = \tau\log\Big(\big\|\exp\big(Q(s,\cdot)/\tau\big)\big\|_1\Big),
\end{aligned} \tag{A.40}$$

where in the second line, equality is attained if $\pi(\cdot|s) \propto \exp(Q^{(()}s,\cdot)/\tau)$. This immediately gives rise to

$$\begin{aligned}
\mathcal{T}_\tau(Q)(s,a) &= r(s,a) + \gamma \mathop{\mathbb{E}}_{s'\sim P(\cdot|s,a)}\left[\max_{\pi(\cdot|s')\in\Delta(\mathcal{A})} \mathop{\mathbb{E}}_{a'\sim\pi(\cdot|s')}\Big[Q(s',a') - \tau\log\pi(a'|s')\Big]\right] \\
&= r(s,a) + \gamma \mathop{\mathbb{E}}_{s'\sim P(\cdot|s,a)}\Big[\tau\log\Big(\big\|\exp\big(Q^{(()}s',\cdot)/\tau\big)\big\|_1\Big)\Big].
\end{aligned}$$

Proof of Eqn. (A.6). Recall the characterization of $\pi_\tau^\star$ and $V_\tau^\star$ established in Nachum et al. [2017]:

$$\pi_\tau^\star(a|s) = \exp\left(\frac{Q_\tau^\star(s,a) - V_\tau^\star(s)}{\tau}\right), \tag{A.41a}$$

$$V_\tau^\star(s) = \tau\log\Big(\big\|\exp\big(Q_\tau^\star(s,\cdot)/\tau\big)\big\|_1\Big). \tag{A.41b}$$

Substitution into the expression (A.5) tells us that for any $(s,a) \in \mathcal{S}\times\mathcal{A}$,

$$\begin{aligned}
\mathcal{T}_\tau\big(Q_\tau^\star\big)(s,a) &= r(s,a) + \gamma \mathop{\mathbb{E}}_{s'\sim P(\cdot|s,a)}\Big[\tau\log\Big(\big\|\exp\big(Q_\tau^\star(s',\cdot)/\tau\big)\big\|_1\Big)\Big] \\
&= r(s,a) + \gamma \mathop{\mathbb{E}}_{s'\sim P(\cdot|s,a)}\big[V_\tau^\star(s')\big] \\
&= Q_\tau^\star(s,a),
\end{aligned}$$

where the second line results from (A.41b), and the last line follows from the definition of the soft Q-function.

Proof of Eqn. (A.7). Invoking again the expression (A.5), we can demonstrate that for any Q_1 and Q_2,

$$\begin{aligned}
&\Big|\mathcal{T}_\tau(Q_1)(s,a) - \mathcal{T}_\tau(Q_2)(s,a)\Big| \\
&\quad= \left|\gamma \mathop{\mathbb{E}}_{s'\sim P(\cdot|s,a)}\Big[\tau\log\Big(\big\|\exp\big(Q_1(s',\cdot)/\tau\big)\big\|_1\Big)\Big] - \gamma \mathop{\mathbb{E}}_{s'\sim P(\cdot|s,a)}\Big[\tau\log\Big(\big\|\exp\big(Q_2(s',\cdot)/\tau\big)\big\|_1\Big)\Big]\right| \\
&\quad= \gamma\tau\left|\mathop{\mathbb{E}}_{s'\sim P(\cdot|s,a)}\Big[\log\Big(\big\|\exp\big(Q_1(s',\cdot)/\tau\big)\big\|_1\Big) - \log\Big(\big\|\exp\big(Q_2(s',\cdot)/\tau\big)\big\|_1\Big)\Big]\right| \\
&\quad\le \gamma\tau\,\big\|Q_1/\tau - Q_2/\tau\big\|_\infty = \gamma\,\big\|Q_1 - Q_2\big\|_\infty
\end{aligned}$$

holds for all $(s,a) \in \mathcal{S} \times \mathcal{A}$, where the inequality follows from the Lipschitz property (A.33).

A.3.3 Proof of Lemma 5

For any state-action pair $(s,a) \in \mathcal{S} \times \mathcal{A}$, we observe that

$$\begin{aligned} &Q_\tau^\star(s,a) - Q_\tau^{(t+1)}(s,a) \\ &\quad = r(s,a) + \gamma \mathop{\mathbb{E}}_{s' \sim P(\cdot|s,a)} \left[V_\tau^\star(s')\right] - \left(r(s,a) + \gamma \mathop{\mathbb{E}}_{s' \sim P(\cdot|s,a)} \left[V_\tau^{(t+1)}(s')\right]\right) \\ &\quad = \gamma \mathop{\mathbb{E}}_{s' \sim P(\cdot|s,a)} \left[\tau \log \left(\left\| \exp\left(\frac{Q_\tau^\star(s',\cdot)}{\tau}\right)\right\|_1\right)\right] - \gamma \mathop{\mathbb{E}}_{\substack{s' \sim P(\cdot|s,a), \\ a' \sim \pi^{(t+1)}(\cdot|s')}} \left[Q_\tau^{(t+1)}(s',a') - \tau \log \pi^{(t+1)}(a'|s')\right], \end{aligned} \tag{A.42}$$

where the first step invokes the definition (2.10a) of Q_τ, and the second step is due to the expression (A.41b) of $V_\tau^\star$. To continue, recall that $\pi^{(t)}$ is related to $\xi^{(t)}$ as

$$\forall s \in \mathcal{S}: \qquad \pi^{(t)}(\cdot|s) = \frac{1}{\|\xi^{(t)}(s,\cdot)\|_1} \xi^{(t)}(s,\cdot) \tag{A.43}$$

which can be seen by comparing (A.11) with (2.17). This in turn leads to

$$\begin{aligned} \log \pi^{(t+1)}(a|s) &= \log \xi^{(t+1)}(s,a) - \log\left(\left\|\xi^{(t+1)}(s,\cdot)\right\|_1\right) \\ &= \alpha \log \xi^{(t)}(s,a) + (1-\alpha)\frac{Q_\tau^{(t)}(s,a)}{\tau} - \log\left(\left\|\xi^{(t+1)}(s,\cdot)\right\|_1\right), \end{aligned} \tag{A.44}$$

where the second line comes from (A.11b). By plugging (A.44) into (A.42) we obtain

$$\begin{aligned} &Q_\tau^\star(s,a) - Q_\tau^{(t+1)}(s,a) \\ &= \gamma \mathop{\mathbb{E}}_{s' \sim P(\cdot|s,a)} \left[\tau \log\left(\left\|\exp\left(Q_\tau^\star(s',\cdot)/\tau\right)\right\|_1\right) - \tau \log\left(\left\|\xi^{(t+1)}(s',\cdot)\right\|_1\right)\right] \\ &\quad - \gamma \mathop{\mathbb{E}}_{\substack{s' \sim P(\cdot|s,a), \\ a' \sim \pi^{(t+1)}(\cdot|s')}} \left[Q_\tau^{(t+1)}(s',a') - \tau \underbrace{\left(\alpha \log \xi^{(t)}(s',a') + (1-\alpha)\frac{Q_\tau^{(t)}(s',a')}{\tau}\right)}_{= \log \xi^{(t+1)}(s',a')}\right] \end{aligned} \tag{A.45}$$

for any $(s,a) \in \mathcal{S} \times \mathcal{A}$. In the sequel, we bound each term on the right-hand side of (A.45) separately.

- In view of the property (A.33), the first term on the right-hand side of (A.45) can be bounded by

$$\tau \log\left(\left\|\exp\left(Q_\tau^\star(s',\cdot)/\tau\right)\right\|_1\right) - \tau \log\left(\left\|\xi^{(t+1)}(s',\cdot)\right\|_1\right) \le \left\|Q_\tau^\star - \tau \log \xi^{(t+1)}\right\|_\infty.$$

- Regarding the second term, the monotonicity (A.2) of the soft Q-function allows us to derive

$$\begin{aligned}
&Q_\tau^{(t+1)}(s,a) - \tau\left(\alpha\log\xi^{(t)}(s,a) + (1-\alpha)Q_\tau^{(t)}(s,a)/\tau\right) \\
&\quad\geq Q_\tau^{(t)}(s,a) - \tau\left(\alpha\log\xi^{(t)}(s,a) + (1-\alpha)Q_\tau^{(t)}(s,a)/\tau\right) \\
&\quad= \alpha\left(Q_\tau^{(t)}(s,a) - \tau\log\xi^{(t)}(s,a)\right) \\
&\quad\overset{\text{(i)}}{=} \alpha\left(\alpha\left(Q_\tau^{(t-1)}(s,a) - \tau\log\xi^{(t-1)}(s,a)\right) + Q_\tau^{(t)}(s,a) - Q_\tau^{(t-1)}(s,a)\right) \\
&\quad\overset{\text{(ii)}}{\geq} \alpha^2\left(Q_\tau^{(t-1)}(s,a) - \tau\log\xi^{(t-1)}(s,a)\right) \\
&\quad\overset{\text{(iii)}}{\geq} \alpha^{t+1}\left(Q_\tau^{(0)}(s,a) - \tau\log\xi^{(0)}(s,a)\right) \\
&\quad\overset{\text{(iv)}}{\geq} -\alpha^{t+1}\big\|Q_\tau^{(0)} - \tau\log\xi^{(0)}\big\|_\infty
\end{aligned}$$

  for any $(s,a)\in\mathcal{S}\times\mathcal{A}$. Here, (i) follows by construction (A.11b), (ii) invokes the monotonicity property (A.2) (so that $Q_\tau^{(t)}\geq Q_\tau^{(t-1)}$), and (iii) follows by repeating the arguments (i) and (ii) recursively.

Combining the preceding two bounds with the expression (A.45), we conclude that

$$0\leq Q_\tau^\star(s,a) - Q_\tau^{(t+1)}(s,a)\leq\gamma\big\|Q_\tau^\star - \tau\log\xi^{(t+1)}\big\|_\infty + \gamma\alpha^{t+1}\big\|Q_\tau^{(0)} - \tau\log\xi^{(0)}\big\|_\infty \tag{A.46}$$

for any $(s,a)\in\mathcal{S}\times\mathcal{A}$, thus concluding the proof.

A.3.4 Proof of Lemma 6

Recall that, in this scenario, the policies are updated using inexact policy evaluation via (2.22), namely,

$$\forall(s,a)\in\mathcal{S}\times\mathcal{A},\qquad \pi^{(t+1)}(a|s) = \frac{\big(\pi^{(t)}(a|s)\big)^{1-\frac{\eta\tau}{1-\gamma}}\exp\big(\frac{\eta}{1-\gamma}\widehat{Q}_\tau^{(t)}(s,a)\big)}{\widehat{Z}^{(t)}(s)}, \tag{A.47}$$

where $\widehat{Z}^{(t)}(s) := \sum_{a'}\pi^{(t)}(a'|s)^{1-\frac{\eta\tau}{1-\gamma}}\exp\big(\frac{\eta}{1-\gamma}\widehat{Q}_\tau^{(t)}(s,a')\big)$. To facilitate analysis, we further introduce another auxiliary policy sequence $\{\breve{\pi}^{(t)}\}$, which corresponds to the policy update as if we had access to exact soft Q-function of $\pi^{(t)}$ in the t-th iteration; this is defined as

$$\forall(s,a)\in\mathcal{S}\times\mathcal{A},\qquad \breve{\pi}^{(t+1)}(a|s) = \frac{\big(\pi^{(t)}(a|s)\big)^{1-\frac{\eta\tau}{1-\gamma}}\exp\big(\frac{\eta}{1-\gamma}Q_\tau^{(t)}(s,a)\big)}{Z^{(t)}(s)}, \tag{A.48}$$

where we abuse the notation by letting $Z^{(t)}(s) := \sum_{a'}\pi^{(t)}(a'|s)^{1-\frac{\eta\tau}{1-\gamma}}\exp\big(\frac{\eta}{1-\gamma}Q_\tau^{(t)}(s,a')\big)$. It is worth emphasizing that $\breve{\pi}^{(t+1)}$ is produced on the basis of $\pi^{(t)}$ as opposed to $\breve{\pi}^{(t)}$; it should be viewed as a one-step perfect update from a given policy $\pi^{(t)}$.

We first make note of the following fact: for any step size $0<\eta\leq(1-\gamma)/\tau$, it follows from (A.34) — together with the construction (A.47) and (A.48) — that

$$\begin{aligned}
&\big\|\log\pi^{(t+1)} - \log\breve{\pi}^{(t+1)}\big\|_\infty \\
&\leq 2\left\|\log\left(\pi^{(t)}(a|s)^{1-\eta\tau/(1-\gamma)}\exp\left(\frac{\eta}{1-\gamma}\widehat{Q}_\tau^{(t)}(s,a)\right)\right) - \log\left(\pi^{(t)}(a|s)^{1-\eta\tau/(1-\gamma)}\exp\left(\frac{\eta}{1-\gamma}Q_\tau^{(t)}(s,a)\right)\right)\right\|_\infty \\
&= \frac{2\eta}{1-\gamma}\big\|\widehat{Q}_\tau^{(t)} - Q_\tau^{(t)}\big\|_\infty.
\end{aligned} \tag{A.49}$$

Next, let us recall the inequality (A.37) in the proof of Lemma 3 under exact policy evaluation $\breve{\pi}^{(t+1)}(\cdot|s)$; when applied to the current setting, it essentially indicates that

$$\begin{aligned} V_\tau^{(t)}(s_0) &= \mathop{\mathbb{E}}_{a_0\sim\breve{\pi}^{(t+1)}(\cdot|s_0)}\left[\frac{1-\gamma}{\eta}\log Z^{(t)}(s_0)\right] - \frac{1-\gamma}{\eta}\mathsf{KL}\left(\pi^{(t)}(\cdot|s_0)\,\|\,\breve{\pi}^{(t+1)}(\cdot|s_0)\right) \\ &= \mathop{\mathbb{E}}_{a_0\sim\pi^{(t+1)}(\cdot|s_0)}\left[\frac{1-\gamma}{\eta}\log Z^{(t)}(s_0)\right] - \frac{1-\gamma}{\eta}\mathsf{KL}\left(\pi^{(t)}(\cdot|s_0)\,\|\,\breve{\pi}^{(t+1)}(\cdot|s_0)\right), \end{aligned} \tag{A.50}$$

where the last step follows since the quantity $Z^{(t)}(s)$ does not depend on a at all. In order to control the first term of (A.50), we invoke the definition of $\breve{\pi}^{(t+1)}(\cdot|s)$ to show that

$$\begin{aligned} &\mathop{\mathbb{E}}_{a_0\sim\pi^{(t+1)}(\cdot|s_0)}\left[\frac{1-\gamma}{\eta}\log Z^{(t)}(s_0)\right] \\ &\overset{\text{(i)}}{=} \mathop{\mathbb{E}}_{a_0\sim\pi^{(t+1)}(\cdot|s_0)}\left[-\tau\log\breve{\pi}^{(t+1)}(a_0|s_0) + Q_\tau^{(t)}(s_0,a_0) + \left(\tau-\frac{1-\gamma}{\eta}\right)\left(\log\breve{\pi}^{(t+1)}(a_0|s_0) - \log\pi^{(t)}(a_0|s_0)\right)\right] \\ &= \mathop{\mathbb{E}}_{a_0\sim\pi^{(t+1)}(\cdot|s_0)}\left[-\tau\log\pi^{(t+1)}(a_0|s_0) + Q_\tau^{(t)}(s_0,a_0)\right] + \left(\tau-\frac{1-\gamma}{\eta}\right)\mathsf{KL}\left(\pi^{(t+1)}(\cdot|s_0)\,\|\,\pi^{(t)}(\cdot|s_0)\right) \\ &\qquad - \frac{1-\gamma}{\eta}\mathop{\mathbb{E}}_{a_0\sim\pi^{(t+1)}(\cdot|s_0)}\left[\log\breve{\pi}^{(t+1)}(a_0|s_0) - \log\pi^{(t)}(a_0|s_0)\right] \\ &\le \mathop{\mathbb{E}}_{a_0\sim\pi^{(t+1)}(\cdot|s_0)}\left[-\tau\log\pi^{(t+1)}(a_0|s_0) + Q_\tau^{(t)}(s_0,a_0)\right] + \left(\tau-\frac{1-\gamma}{\eta}\right)\mathsf{KL}\left(\pi^{(t+1)}(\cdot|s_0)\,\|\,\pi^{(t)}(\cdot|s_0)\right) \\ &\qquad + 2\big\|\widehat{Q}_\tau^{(t)} - Q_\tau^{(t)}\big\|_\infty, \end{aligned} \tag{A.51}$$

where the final step results from (A.49). Putting the above bound together with (A.50) guarantees that

$$\begin{aligned} V_\tau^{(t)}(s_0) &\le \mathop{\mathbb{E}}_{a_0\sim\pi^{(t+1)}(\cdot|s_0)}\left[-\tau\log\pi^{(t+1)}(a_0|s_0) + Q_\tau^{(t)}(s_0,a_0)\right] - \frac{1-\gamma}{\eta}\mathsf{KL}\left(\pi^{(t)}(\cdot|s_0)\,\|\,\breve{\pi}^{(t+1)}(\cdot|s_0)\right) \\ &\qquad - \left(\frac{1-\gamma}{\eta}-\tau\right)\mathsf{KL}\left(\pi^{(t+1)}(\cdot|s_0)\,\|\,\pi^{(t)}(\cdot|s_0)\right) + 2\big\|\widehat{Q}_\tau^{(t)} - Q_\tau^{(t)}\big\|_\infty \\ &\le \mathop{\mathbb{E}}_{a_0\sim\pi^{(t+1)}(\cdot|s_0)}\left[-\tau\log\pi^{(t+1)}(a_0|s_0) + Q_\tau^{(t)}(s_0,a_0)\right] + 2\big\|\widehat{Q}_\tau^{(t)} - Q_\tau^{(t)}\big\|_\infty \\ &\le \mathop{\mathbb{E}}_{a_0\sim\pi^{(t+1)}(\cdot|s_0)}\left[-\tau\log\pi^{(t+1)}(a_0|s_0) + r(s_0,a_0) + \gamma\mathop{\mathbb{E}}_{s_1\sim P(\cdot|s_0,a_0)}\left[V_\tau^{(t)}(s_1)\right]\right] + 2\big\|\widehat{Q}_\tau^{(t)} - Q_\tau^{(t)}\big\|_\infty. \end{aligned}$$

where the last identity makes use of the relation $Q_\tau^{(t)}(s_0,a_0) = r(s_0,a_0) + \gamma\mathbb{E}_{s_1\sim P(\cdot|s_0,a_0)}\big[V_\tau^{(t)}(s_1)\big]$. Invoking the above inequality recursively as in the expression (A.39) (see Lemma 3), we can expand it to establish

$$V_\tau^{(t)}(s_0) \le V_\tau^{(t+1)}(s_0) + 2\big\|\widehat{Q}_\tau^{(t)} - Q_\tau^{(t)}\big\|_\infty\sum_{i=0}^{\infty}\gamma^i = V_\tau^{(t+1)}(s_0) + \frac{2}{1-\gamma}\big\|\widehat{Q}_\tau^{(t)} - Q_\tau^{(t)}\big\|_\infty.$$

A.3.5 Proof of Lemma 7

The results of this lemma, or some similar versions, have appeared in prior work (e.g. Mei et al. [2020b, Lemma 10] and Agarwal et al. [2020b, Lemma 5.6]). We include the proof here primarily for the sake of self-completeness.

Proof of Eqn. (A.30a). The policy gradient of the unregularized value function $V^{\pi_\theta}(s_0)$ is well-known as the policy gradient theorem [Sutton et al., 2000]. Here, we deal with a slightly different variant – an entropy-regularized value function $V_\tau^{\pi_\theta}(s_0)$ in the expression (2.1) with the softmax policy parameterization in (2.6). Invoking the Bellman equation and recognizing that $V_\tau^{\pi_\theta}(s_0)$ can be viewed as an unregularized value function with instantaneous rewards $r(s,a) - \tau \log \pi_\theta(a|s)$ for any (s,a), we obtain

$$
\begin{aligned}
&\nabla_\theta V_\tau^{\pi_\theta}(s_0) \\
&= \nabla_\theta \left[\sum_{a_0} \pi_\theta(a_0|s_0)\Big(r(s_0,a_0) - \tau\log\pi_\theta(a_0|s_0) + \gamma \mathop{\mathbb{E}}_{s'\sim P(\cdot|s_0,a_0)}\big[V_\tau^{\pi_\theta}(s')\big]\Big)\right] \\
&\overset{\text{(i)}}{=} \nabla_\theta \left[\sum_{a_0} \pi_\theta(a_0|s_0)\Big(Q_\tau^{\pi_\theta}(s_0,a_0) - \tau\log\pi_\theta(a_0|s_0)\Big)\right] \\
&= \sum_{a_0} \big(\nabla_\theta \pi_\theta(a_0|s_0)\big)\Big(Q_\tau^{\pi_\theta}(s_0,a_0) - \tau\log\pi_\theta(a_0|s_0)\Big) + \sum_{a_0}\pi_\theta(a_0|s_0)\nabla_\theta\Big(Q_\tau^{\pi_\theta}(s_0,a_0) - \tau\log\pi_\theta(a_0|s_0)\Big) \\
&\overset{\text{(ii)}}{=} \sum_{a_0} \Big(\pi_\theta(a_0|s_0)\nabla_\theta\log\pi_\theta(a_0|s_0)\Big)\Big(Q_\tau^{\pi_\theta}(s_0,a_0) - \tau\log\pi_\theta(a_0|s_0)\Big) \\
&\qquad + \sum_{a_0}\pi_\theta(a_0|s_0)\nabla_\theta\Big(r(s_0,a_0) + \gamma\sum_{s_1} P(s_1|s_0,a_0)V_\tau^{\pi_\theta}(s_1) - \tau\log\pi_\theta(a_0|s_0)\Big),
\end{aligned}
$$

where (i) relies on the definition (2.10a) of $Q_\tau^{\pi_\theta}$, and (ii) makes use of the identity

$$
\nabla_\theta \pi_\theta(a_0|s_0) = \pi_\theta(a_0|s_0)\nabla_\theta \log \pi_\theta(a_0|s_0)
$$

as well as the definition (2.10a) of $Q_\tau^{\pi_\theta}$. Given that

$$
\sum_{a_0} \pi_\theta(a_0|s_0)\nabla_\theta\log\pi_\theta(a_0|s_0) = \sum_{a_0}\nabla_\theta\pi_\theta(a_0|s_0) = \nabla_\theta\Big(\sum_{a_0}\pi_\theta(a_0|s_0)\Big) = \nabla_\theta 1 = 0 \tag{A.52}
$$

and that $r(s,a)$ is independent of θ, one can continue the above derivative to reach

$$
\begin{aligned}
\nabla_\theta V_\tau^{\pi_\theta}(s_0) &= \sum_{a_0} \Big(\pi_\theta(a_0|s_0)\nabla_\theta\log\pi_\theta(a_0|s_0)\Big)\Big(Q_\tau^{\pi_\theta}(s_0,a_0) - \tau\log\pi_\theta(a_0|s_0)\Big) \\
&\qquad + \gamma\sum_{a_0}\pi_\theta(a_0|s_0)\sum_{s_1}P(s_1|s_0,a_0)\nabla_\theta V_\tau^{\pi_\theta}(s_1) \\
&= \mathop{\mathbb{E}}_{\substack{a_i\sim\pi_\theta(\cdot|s_i),\\ s_{i+1}\sim P(\cdot|s_i,a_i),\forall i\geq 0}} \Big[\big(\nabla_\theta\log\pi_\theta(a_0|s_0)\big)\Big(Q_\tau^{\pi_\theta}(s_0,a_0) - \tau\log\pi_\theta(a_0|s_0)\Big) + \gamma\nabla_\theta V_\tau^{\pi_\theta}(s_1)\Big].
\end{aligned}
$$

Repeating the above calculations recursively, we arrive at

$$
\begin{aligned}
\nabla_\theta V_\tau^{\pi_\theta}(s_0) &= \mathop{\mathbb{E}}_{\substack{a_i\sim\pi_\theta(\cdot|s_i),\\ s_{i+1}\sim P(\cdot|s_i,a_i),\forall i\geq 0}} \left[\sum_{t=0}^{\infty}\gamma^t\big(\nabla_\theta\log\pi_\theta(a_t|s_t)\big)\Big(Q_\tau^{\pi_\theta}(s_t,a_t) - \tau\log\pi_\theta(a_t|s_t)\Big)\right] \\
&= \frac{1}{1-\gamma}\mathop{\mathbb{E}}_{s\sim d_{s_0}^{\pi_\theta}}\mathop{\mathbb{E}}_{a\sim\pi_\theta(\cdot|s)}\Big[\big(\nabla_\theta\log\pi_\theta(a|s)\big)\Big(Q_\tau^{\pi_\theta}(s,a) - \tau\log\pi_\theta(a|s)\Big)\Big] \\
&= \frac{1}{1-\gamma}\mathop{\mathbb{E}}_{s\sim d_{s_0}^{\pi_\theta}}\mathop{\mathbb{E}}_{a\sim\pi_\theta(\cdot|s)}\Big[\big(\nabla_\theta\log\pi_\theta(a|s)\big)\Big(A_\tau^{\pi_\theta}(s,a) + V_\tau^{\pi_\theta}(s)\Big)\Big] \\
&= \frac{1}{1-\gamma}\mathop{\mathbb{E}}_{s\sim d_{s_0}^{\pi_\theta}}\mathop{\mathbb{E}}_{a\sim\pi_\theta(\cdot|s)}\Big[\big(\nabla_\theta\log\pi_\theta(a|s)\big)A_\tau^{\pi_\theta}(s,a)\Big],
\end{aligned} \tag{A.53}
$$

where the second line follows by aggregating the terms corresponding to the same state-action pair, and the third line invokes the definition (A.29) of $A_\tau^{\pi_\theta}$. To see why the last line holds, invoke (A.52) to reach

$$\begin{aligned}\mathbb{E}_{a\sim\pi_\theta(\cdot|s)}\Big[V_\tau^{\pi_\theta}(s)\nabla_\theta\log\pi_\theta(a|s)\Big] &= \sum_a V_\tau^{\pi_\theta}(s)\pi_\theta(a|s)\nabla_\theta\log\pi_\theta(a|s)\\ &= V_\tau^{\pi_\theta}(s)\sum_a \pi_\theta(a|s)\nabla_\theta\log\pi_\theta(a|s) = 0.\end{aligned}$$

Further, it is easily seen that under the softmax parametrization in (2.6),

$$\frac{\partial\log\pi_\theta(a'|s')}{\partial\theta(s,a)} = \mathbb{1}[s'=s]\big(\mathbb{1}[a'=a]-\pi_\theta(a|s)\big) \tag{A.54}$$

for any $(s,a),(s',a')\in\mathcal{S}\times\mathcal{A}$. Combining with (A.53), it further implies that

$$\begin{aligned}\frac{\partial V_\tau^{\pi_\theta}(s_0)}{\partial\theta(s,a)} &= \frac{1}{1-\gamma}\mathbb{E}_{s'\sim d_{s_0}^{\pi_\theta}}\mathbb{E}_{a'\sim\pi_\theta(\cdot|s')}\left[\frac{\partial\log\pi_\theta(a'|s')}{\partial\theta(s,a)}A_\tau^{\pi_\theta}(s',a')\right]\\ &= \frac{1}{1-\gamma}\mathbb{E}_{s'\sim d_{s_0}^{\pi_\theta}}\mathbb{E}_{a'\sim\pi_\theta(\cdot|s')}\left[\Big(\mathbb{1}[s'=s]\big(\mathbb{1}[a'=a]-\pi_\theta(a|s)\big)\Big)A_\tau^{\pi_\theta}(s',a')\right]\\ &\overset{\text{(i)}}{=} \frac{1}{1-\gamma}\mathbb{E}_{s'\sim d_{s_0}^{\pi_\theta}}\mathbb{E}_{a'\sim\pi_\theta(\cdot|s')}\Big[\mathbb{1}\left[(s',a')=(s,a)\right]A_\tau^{\pi_\theta}(s',a')\Big]\\ &= \frac{1}{1-\gamma}d_{s_0}^{\pi_\theta}(s)\pi_\theta(a|s)A_\tau^{\pi_\theta}(s,a).\end{aligned}$$

where (i) follows from $\mathbb{E}_{a'\sim\pi_\theta(\cdot|s')}A_\tau^{\pi_\theta}(s',a') = \sum_{a'}\pi_\theta(a'|s')A_\tau^{\pi_\theta}(s',a') = 0$ due to the definition (A.29). The proof regarding $V_\tau^{\pi_\theta}(\rho)$ can be obtained by averaging the initial state s_0 over the distribution ρ.

Proof of Eqn. (A.30b). In order to establish (A.30b), a crucial observation is that $w_\theta := \big(\mathcal{F}_\rho^\theta\big)^\dagger\nabla_\theta V_\tau^{\pi_\theta}(\rho)$ is exactly the solution to the following least-squares problem

$$\text{minimize}_{w\in\mathbb{R}^{|\mathcal{S}||\mathcal{A}|}}\quad \big\|\mathcal{F}_\rho^\theta w - \nabla_\theta V_\tau^{\pi_\theta}(\rho)\big\|_2^2. \tag{A.55}$$

From the definition (2.13) of the Fisher information matrix, we have

$$\mathcal{F}_\rho^\theta w = \mathbb{E}_{s\sim d_\rho^{\pi_\theta}}\mathbb{E}_{a\sim\pi_\theta(\cdot|s)}\Big[\big(\nabla_\theta\log\pi_\theta(a|s)\big)\big(\nabla_\theta\log\pi_\theta(a|s)\big)^\top w\Big].$$

for any fixed vector $w = [w_{s,a}]_{(s,a)\in\mathcal{S}\times\mathcal{A}}$. As a result, for any $(s,a)\in\mathcal{S}\times\mathcal{A}$ one has

$$
\begin{aligned}
&\Big(\mathcal{F}_\rho^\theta w\Big)_{s,a} \\
&= \mathbb{E}_{s'\sim d_\rho^{\pi_\theta}} \mathbb{E}_{a'\sim\pi_\theta(\cdot|s')}\left[\frac{\partial \log \pi_\theta(a'|s')}{\partial\theta(s,a)}\left(\sum_{\tilde{s},\tilde{a}} \frac{\partial \log \pi_\theta(a'|s')}{\partial\theta(\tilde{s},\tilde{a})} w_{\tilde{s},\tilde{a}}\right)\right] \\
&\overset{\text{(i)}}{=} \mathbb{E}_{s'\sim d_\rho^{\pi_\theta}} \mathbb{E}_{a'\sim\pi_\theta(\cdot|s')}\left[\mathbb{1}[s'=s]\Big(\mathbb{1}[a'=a]-\pi_\theta(a|s)\Big)\left(\sum_{\tilde{s},\tilde{a}} \mathbb{1}[\tilde{s}=s']\Big(\mathbb{1}[\tilde{a}=a']-\pi_\theta(\tilde{a}|\tilde{s})\Big) w_{\tilde{s},\tilde{a}}\right)\right] \\
&= \mathbb{E}_{s'\sim d_\rho^{\pi_\theta}} \mathbb{E}_{a'\sim\pi_\theta(\cdot|s')}\left[\mathbb{1}[s'=s]\Big(\mathbb{1}[a'=a]-\pi_\theta(a|s)\Big)\Big(w_{s',a'}-\sum_{\tilde{a}}\pi_\theta(\tilde{a}|s')w_{s',\tilde{a}}\Big)\right] \\
&= d_\rho^{\pi_\theta}(s)\mathbb{E}_{a'\sim\pi_\theta(\cdot|s')}\left[\Big(\mathbb{1}[a'=a]-\pi_\theta(a|s)\Big)\big(w_{s,a'}-c(s)\big)\right] \\
&= d_\rho^{\pi_\theta}(s)\mathbb{E}_{a'\sim\pi_\theta(\cdot|s')}\Big[\mathbb{1}[a'=a]w_{s,a'}-\pi_\theta(a|s)w_{s,a'}-\mathbb{1}[a'=a]c(s)+\pi_\theta(a|s)c(s)\Big] \\
&= d_\rho^{\pi_\theta}(s)\Big[\pi_\theta(a|s)w_{s,a}-\pi_\theta(a|s)c(s)-\pi_\theta(a|s)c(s)+\pi_\theta(a|s)c(s)\Big] \\
&= d_\rho^{\pi_\theta}(s)\pi_\theta(a|s)\big[w_{s,a}-c(s)\big],
\end{aligned}
$$

where (i) makes use of the derivative calculation (A.54), and we define $c(s) := \sum_a \pi_\theta(a|s)w_{s,a}$. Consequently, the objective function of (A.55) can be written as

$$
\begin{aligned}
\big\|\mathcal{F}_\rho^\theta w - \nabla_\theta V_\tau^{\pi_\theta}(\rho)\big\|_2^2 &= \sum_{s,a}\left(d_\rho^{\pi_\theta}(s)\pi_\theta(a|s)\left[w_{s,a}-c(s)\right]-\frac{1}{1-\gamma}d_\rho^{\pi_\theta}(s)\pi_\theta(a|s)A_\tau^{\pi_\theta}(s,a)\right)^2 \\
&= \sum_{s,a}\left(d_\rho^{\pi_\theta}(s)\pi_\theta(a|s)\left(w_{s,a}-c(s)-\frac{1}{1-\gamma}A_\tau^{\pi_\theta}(s,a)\right)\right)^2,
\end{aligned}
$$

which is minimized by choosing $w_{s,a} = \frac{1}{1-\gamma}A_\tau^{\pi_\theta}(s,a)+c(s)$ for all $(s,a)\in\mathcal{S}\times\mathcal{A}$. This concludes the proof.

A.4 Proof for approximate entropy-regularized NPG (Theorem 2)

In this section, we complete the proofs of Theorem 2 in Section A.1.3, which consists of (i) establishing the linear system in (A.27) and (ii) extracting the convergence rate from (A.27).

Step 1: establishing the linear system (A.27). In what follows, we shall justify the linear system relation by checking each row separately.

(1) **Bounding $\|Q_\tau^\star - \tau\log\widehat{\xi}^{(t+1)}\|_\infty$.** From the construction (A.26b) of $\widehat{\xi}^{(t+1)}$, we have

$$
Q_\tau^\star - \tau\log\widehat{\xi}^{(t+1)} = \alpha\big(Q_\tau^\star - \tau\log\widehat{\xi}^{(t)}\big) + (1-\alpha)\big(Q_\tau^\star - Q_\tau^{(t)}\big) + (1-\alpha)\big(Q_\tau^{(t)} - \widehat{Q}_\tau^{(t)}\big).
$$

Taken together with the triangle inequality and the assumption $\big\|Q_\tau^{(t)} - \widehat{Q}_\tau^{(t)}\big\|_\infty \le \delta$, this gives

$$
\big\|Q_\tau^\star - \tau\log\widehat{\xi}^{(t+1)}\big\|_\infty \le \alpha\big\|Q_\tau^\star - \tau\log\widehat{\xi}^{(t)}\big\|_\infty + (1-\alpha)\big\|Q_\tau^\star - Q_\tau^{(t)}\big\|_\infty + (1-\alpha)\,\delta. \tag{A.56}
$$

(2) **Bounding** $-\min_{s,a}\left(Q_\tau^{(t+1)}(s,a)-\tau\log\widehat{\xi}^{(t+1)}(s,a)\right)$**.** Invoking the definition (A.26b) of $\widehat{\xi}^{(t+1)}$ again implies that for any $(s,a)\in\mathcal{S}\times\mathcal{A}$,

$$
\begin{aligned}
&-\left(Q_\tau^{(t+1)}(s,a)-\tau\log\widehat{\xi}^{(t+1)}(s,a)\right)\\
&\quad=-\left(Q_\tau^{(t+1)}(s,a)-\tau\left(\alpha\log\widehat{\xi}^{(t)}(s,a)+(1-\alpha)\widehat{Q}_\tau^{(t)}(s,a)/\tau\right)\right)\\
&\quad=-\alpha\left(Q_\tau^{(t)}(s,a)-\tau\log\widehat{\xi}^{(t)}(s,a)\right)+(1-\alpha)\left(\widehat{Q}_\tau^{(t)}(s,a)-Q_\tau^{(t)}(s,a)\right)+\left(Q_\tau^{(t)}(s,a)-Q_\tau^{(t+1)}(s,a)\right)\\
&\quad\le-\alpha\left(Q_\tau^{(t)}(s,a)-\tau\log\widehat{\xi}^{(t)}(s,a)\right)+(1-\alpha)\,\delta+\frac{2\gamma\delta}{1-\gamma},
\end{aligned}
$$

where the last inequality follows from $\left\|Q_\tau^{(t)}-\widehat{Q}_\tau^{(t)}\right\|_\infty\le\delta$ and (A.25). Taking the maximum over $(s,a)\in\mathcal{S}\times\mathcal{A}$ on both sides and using the definition $\alpha=1-\frac{\eta\tau}{1-\gamma}$ yield

$$
-\min_{s,a}\left(Q_\tau^{(t+1)}(s,a)-\tau\log\widehat{\xi}^{(t+1)}(s,a)\right)\le-\alpha\min_{s,a}\left(Q_\tau^{(t)}(s,a)-\tau\log\widehat{\xi}^{(t)}(s,a)\right)+(1-\alpha)\,\delta\left(1+\frac{2\gamma}{\eta\tau}\right). \tag{A.57}
$$

(3) **Bounding** $\left\|Q_\tau^\star-Q_\tau^{(t+1)}\right\|_\infty$**.** Following the same arguments as for (A.45), we obtain

$$
\begin{aligned}
Q_\tau^\star(s,a)-Q_\tau^{(t+1)}(s,a)&=\gamma\mathop{\mathbb{E}}_{s'\sim P(\cdot|s,a)}\left[\tau\log\left(\left\|\exp\left(Q_\tau^\star(s',\cdot)/\tau\right)\right\|_1\right)-\tau\log\left(\left\|\widehat{\xi}^{(t+1)}(s',\cdot)\right\|_1\right)\right]\\
&\quad-\gamma\mathop{\mathbb{E}}_{\substack{s'\sim P(\cdot|s,a),\\ a'\sim\pi^{(t+1)}(\cdot|s')}}\left[Q_\tau^{(t+1)}(s',a')-\tau\log\widehat{\xi}^{(t+1)}(s',a')\right]\\
&\le\gamma\left\|Q_\tau^\star-\tau\log\widehat{\xi}^{(t+1)}\right\|_\infty-\gamma\min_{s,a}\left(Q_\tau^{(t+1)}(s,a)-\tau\log\widehat{\xi}^{(t+1)}(s,a)\right),
\end{aligned}
$$

where the last line follows from (A.33). By plugging (A.56) and (A.57) into the above inequality, we arrive at the claimed bound regarding this term.

Step 2: deducing convergence guarantees from the linear system (A.27)**.** We start by pinning down the eigenvalues and eigenvectors of the matrix B. Specifically, the three eigenvalues can be calculated as

$$
\lambda_1=\alpha+\gamma(1-\alpha)=1-\eta\tau,\qquad\lambda_2=\alpha\qquad\text{and}\qquad\lambda_3=0, \tag{A.58}
$$

whose corresponding eigenvectors are given respectively by

$$
v_1=\begin{bmatrix}\gamma\\1\\0\end{bmatrix},\qquad v_2=\begin{bmatrix}0\\-1\\1\end{bmatrix},\qquad\text{and}\qquad v_3=\begin{bmatrix}\alpha\\\alpha-1\\0\end{bmatrix}. \tag{A.59}
$$

With some elementary computation, one can show that z_0 and b introduced in (A.28) can be related to the eigenvectors of B in the following way:

$$\begin{aligned}
z_0 &\leq \begin{bmatrix} \left\|Q_\tau^\star - Q_\tau^{(0)}\right\|_\infty \\ \left\|Q_\tau^\star - \tau\log\widehat{\xi}^{(0)}\right\|_\infty \\ \left\|Q_\tau^{(0)} - \tau\log\widehat{\xi}^{(0)}\right\|_\infty \end{bmatrix} \\
&= \frac{1}{1-\eta\tau}\left[(1-\alpha)\left\|Q_\tau^\star - Q_\tau^{(0)}\right\|_\infty + \alpha\left(\left\|Q_\tau^\star - \tau\log\widehat{\xi}^{(0)}\right\|_\infty + \left\|Q_\tau^{(0)} - \tau\log\widehat{\xi}^{(0)}\right\|_\infty\right)\right] v_1 \\
&\qquad + \left\|Q_\tau^{(0)} - \tau\log\widehat{\xi}^{(0)}\right\|_\infty v_2 + c_z v_3 \\
&\leq \frac{1}{1-\eta\tau}\left(\left\|Q_\tau^\star - Q_\tau^{(0)}\right\|_\infty + 2\alpha\tau\left\|\log\pi_\tau^\star - \log\pi^{(0)}\right\|_\infty\right) v_1 + \left\|Q_\tau^{(0)} - \tau\log\widehat{\xi}^{(0)}\right\|_\infty v_2 + c_z v_3,
\end{aligned} \tag{A.60}$$

where c_z is some scalar whose value is immaterial since the eigenvalue corresponding to v_3 is $\lambda_3 = 0$, and the last line follows from the same reasoning for (A.21). Another userful identity is:

$$b = (1-\alpha)\delta\begin{bmatrix} \gamma\left(2+\frac{2\gamma}{\eta\tau}\right) \\ 1 \\ 1+\frac{2\gamma}{\eta\tau}\end{bmatrix} = (1-\alpha)\delta\left[\left(2+\frac{2\gamma}{\eta\tau}\right)v_1 + \left(1+\frac{2\gamma}{\eta\tau}\right)v_2\right]. \tag{A.61}$$

With these preparations in place, we can now invoke the recursion relationship (A.27) and the non-negativity of B to obtain

$$\begin{aligned}
z_{t+1} &\leq B^{t+1}z_0 + \sum_{s=0}^{t} B^{t-s}b \\
&\leq B^{t+1}\left[\frac{1}{1-\eta\tau}\left(\left\|Q_\tau^\star - Q_\tau^{(0)}\right\|_\infty + 2\alpha\tau\left\|\log\pi_\tau^\star - \log\pi^{(0)}\right\|_\infty\right)v_1 + \left\|Q_\tau^{(0)} - \tau\log\widehat{\xi}^{(0)}\right\|_\infty v_2 + c_z v_3\right] \\
&\qquad + (1-\alpha)\delta\sum_{s=0}^{t} B^{t-s}\left[\left(2+\frac{2\gamma}{\eta\tau}\right)v_1 + \left(1+\frac{2\gamma}{\eta\tau}\right)v_2\right] \\
&= \left[\lambda_1^t\left(\left\|Q_\tau^\star - Q_\tau^{(0)}\right\|_\infty + 2\alpha\tau\left\|\log\pi_\tau^\star - \log\pi^{(0)}\right\|_\infty\right) + (1-\alpha)\delta\left(2+\frac{2\gamma}{\eta\tau}\right)\frac{1-\lambda_1^{t+1}}{1-\lambda_1}\right]v_1 \\
&\qquad + \left[\lambda_2^{t+1}\left\|Q_\tau^{(0)} - \tau\log\widehat{\xi}^{(0)}\right\|_\infty + (1-\alpha)\delta\left(1+\frac{2\gamma}{\eta\tau}\right)\frac{1-\lambda_2^{t+1}}{1-\lambda_2}\right]v_2,
\end{aligned}$$

where the eigenvalues and eigenvectors of B are given in (A.58) and (A.59), respectively, and the second inequality relies on (A.60) and (A.61). Note that we are only interested in the first two entries of the vector z_t. Since the first two entries of the eigenvector v_2 are non-positive, we can

safely drop the term involving v_2 in the above inequality to obtain

$$\begin{aligned}
&\begin{bmatrix} \left\|Q_\tau^\star - Q_\tau^{(t+1)}\right\|_\infty \\ \left\|Q_\tau^\star - \tau \log \widehat{\xi}^{(t+1)}\right\|_\infty \end{bmatrix} \\
&\le \left\{ \lambda_1^t \left(\left\|Q_\tau^\star - Q_\tau^{(0)}\right\|_\infty + 2\alpha\tau \left\|\log \pi_\tau^\star - \log \pi^{(0)}\right\|_\infty \right) + (1-\alpha)\delta \left(2 + \frac{2\gamma}{\eta\tau}\right) \frac{1-\lambda_1^{t+1}}{1-\lambda_1} \right\} \begin{bmatrix} \gamma \\ 1 \end{bmatrix} \\
&\le \left\{ (1-\eta\tau)^t \left(\left\|Q_\tau^\star - Q_\tau^{(0)}\right\|_\infty + 2\left(1 - \frac{\eta\tau}{1-\gamma}\right)\tau \left\|\log \pi_\tau^\star - \log \pi^{(0)}\right\|_\infty \right) + \frac{2\delta}{1-\gamma}\left(1 + \frac{\gamma}{\eta\tau}\right) \right\} \begin{bmatrix} \gamma \\ 1 \end{bmatrix}.
\end{aligned} \tag{A.62}$$

When it comes to the log policies, we recall again the fact that $\pi^{(t)}$ is related to $\widehat{\xi}^{(t)}$ as

$$\forall s \in \mathcal{S}: \qquad \pi^{(t)}(\cdot|s) = \frac{1}{\left\|\widehat{\xi}^{(t)}(s,\cdot)\right\|_1} \widehat{\xi}^{(t)}(s,\cdot). \tag{A.63}$$

Invoking the elementary property (A.34), we reach

$$\left\|\log \pi_\tau^\star - \log \pi^{(t+1)}\right\|_\infty \le 2\left\|Q_\tau^\star/\tau - \log \widehat{\xi}^{(t+1)}\right\|_\infty.$$

This together with the bound on $\left\|Q_\tau^\star - \tau \log \widehat{\xi}^{(t+1)}\right\|_\infty$ in (A.62) establishes our claim for $\left\|\log \pi_\tau^\star - \log \pi^{(t+1)}\right\|_\infty$.

Appendix B

Proofs for Chapter 3

B.1 Analysis for exact GPMD (Theorem 3)

In this section, we present the analysis for our main result in Theorem 3, which follows a different framework from Lan [2023]. Here and throughout, we shall often employ the following shorthand notation when it is clear from the context:

$$\begin{aligned} \pi^{(k)}(s) &\coloneqq \pi^{(k)}(\cdot \,|\, s) \in \Delta(\mathcal{A}), \qquad & Q^{\pi}(s) &\coloneqq Q^{\pi}(s,\cdot) \in \mathbb{R}^{|\mathcal{A}|}, \\ \xi^{(k)}(s) &\coloneqq \xi^{(k)}(s,\cdot) \in \mathbb{R}^{|\mathcal{A}|}, & Q_{\tau}^{\pi}(s) &\coloneqq Q_{\tau}^{\pi}(s,\cdot) \in \mathbb{R}^{|\mathcal{A}|}, \end{aligned} \tag{B.1}$$

in addition to those already defined in (3.7).

B.1.1 Preparation: basic facts

In this subsection, we single out a few basic results that underlie the proof of our main theorems.

Performance improvement. To begin with, we demonstrate that GPMD enjoys a sort of monotonic improvements concerning the updates of both the value function and the Q-function, as stated in the following lemma. This lemma can be viewed as a generalization of the well-established policy improvement lemma in the analysis of NPG [Agarwal et al., 2020b, Cen et al., 2022b] as well as PMD [Lan, 2023].

Lemma 8 (Pointwise monotonicity). *For any $(s,a) \in \mathcal{S} \times \mathcal{A}$ and any $k \geq 0$, Algorithm 2 achieves*

$$V_{\tau}^{(k+1)}(s) \geq V_{\tau}^{(k)}(s) \qquad \text{and} \qquad Q_{\tau}^{(k+1)}(s,a) \geq Q_{\tau}^{(k)}(s,a). \tag{B.2}$$

Proof. See Appendix B.2.2. □

Interestingly, the above monotonicity holds simultaneously for all state-action pairs, and hence can be understood as a kind of pointwise monotonicity.

Generalized Bellman operator. Another key ingredient of our proof lies in the use of a generalized Bellman operator $\mathcal{T}_{\tau,h} : \mathbb{R}^{|\mathcal{S}||\mathcal{A}|} \to \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ associated with the regularizer $h = \{h_s\}_{s\in\mathcal{S}}$. Specifically, for any state-action pair (s,a) and any vector $Q \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$, we define

$$\mathcal{T}_{\tau,h}(Q)(s,a) = r(s,a) + \gamma \mathop{\mathbb{E}}_{s'\sim P(\cdot|s,a)} \left[\max_{p\in\Delta(\mathcal{A})} \left\{ \langle Q(s'), p\rangle - \tau h_{s'}(p) \right\} \right]. \tag{B.3}$$

It is worth noting that this definition shares similarity with the regularized Bellman operator proposed in Geist et al. [2019], where the operator defined there is targeted at V_τ, while ours is defined w.r.t. Q_τ.

The importance of this generalized Bellman operator is two-fold: it enjoys a desired contraction property, and its fixed point corresponds to the optimal regularized Q-function. These are generalizations of the properties for the classical Bellman operator, and are formally stated in the following lemma. The proof is deferred to Appendix B.2.3.

Lemma 9 (Properties of the generalized Bellman operator)**.** *For any $\tau > 0$, the operator $\mathcal{T}_{\tau,h}$ defined in (B.3) satisfies the following properties:*

- *$\mathcal{T}_{\tau,h}$ is a contraction operator w.r.t. the ℓ_∞ norm, namely, for any $Q_1, Q_2 \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$, one has*
$$\big\|\mathcal{T}_{\tau,h}(Q_1) - \mathcal{T}_{\tau,h}(Q_2)\big\|_\infty \le \gamma \|Q_1 - Q_2\|_\infty. \tag{B.4}$$
- *The optimal regularized Q-function $Q_\tau^\star$ is a fixed point of $\mathcal{T}_{\tau,h}$, that is,*
$$\mathcal{T}_{\tau,h}(Q_\tau^\star) = Q_\tau^\star. \tag{B.5}$$

B.1.2 Proof of Theorem 3

Similar to the proofs for NPG in Appendix A, our proof consists of (i) characterizing the dynamics of ℓ_∞ errors and establishing a connection to a useful linear system with two variables, and (ii) analyzing the dynamics of this linear system directly. In what follows, we elaborate on each of these steps.

Step 1: error contraction and its connection to a linear system. With the assistance of the above preparations, we are ready to elucidate how to characterize the convergence behavior of $\|Q_\tau^\star - Q_\tau^{(k+1)}\|_\infty$. Recalling the update rule of $\xi^{(k+1)}$ (cf. (3.14c)), we can deduce that

$$Q_\tau^\star - \tau\xi^{(k+1)} = \alpha\big(Q_\tau^\star - \tau\xi^{(k)}\big) + (1-\alpha)\big(Q_\tau^\star - Q_\tau^{(k)}\big)$$

with $\alpha = \frac{1}{1+\eta\tau}$, thus indicating that

$$\big\|Q_\tau^\star - \tau\xi^{(k+1)}\big\|_\infty \le \alpha\big\|Q_\tau^\star - \tau\xi^{(k)}\big\|_\infty + (1-\alpha)\big\|Q_\tau^\star - Q_\tau^{(k)}\big\|_\infty. \tag{B.6}$$

Interestingly, there exists an intimate connection between $\|Q_\tau^\star - Q_\tau^{(k+1)}\|_\infty$ and $\|Q_\tau^\star - \tau\xi^{(k+1)}\|_\infty$ that allows us to bound the former term by the latter. This is stated in the following lemma, with the proof postponed to Appendix B.2.4.

Lemma 10. *Set $\alpha = \frac{1}{1+\eta\tau}$. The iterates of Algorithm 2 satisfy*

$$\big\|Q_\tau^\star - Q_\tau^{(k+1)}\big\|_\infty \le \gamma\big\|Q_\tau^\star - \tau\xi^{(k+1)}\big\|_\infty + \gamma\alpha^{k+1}\big\|Q_\tau^{(0)} - \tau\xi^{(0)}\big\|_\infty. \tag{B.7}$$

The above inequalities (B.6) and (B.7) can be succinctly described via a useful linear system with two variables $\|Q_\tau^\star - Q_\tau^{(k)}\|_\infty$ and $\|Q_\tau^\star - \tau\xi^{(k)}\|_\infty$, that is,

$$x_{k+1} \le Ax_k + \gamma\alpha^{k+1}y, \tag{B.8}$$

where

$$A := \begin{bmatrix}\gamma(1-\alpha) & \gamma\alpha \\ 1-\alpha & \alpha\end{bmatrix}, \qquad x_k := \begin{bmatrix}\|Q_\tau^\star - Q_\tau^{(k)}\|_\infty \\ \|Q_\tau^\star - \tau\xi^{(k)}\|_\infty\end{bmatrix} \qquad \text{and} \qquad y := \begin{bmatrix}\|Q_\tau^{(0)} - \tau\xi^{(0)}\|_\infty \\ 0\end{bmatrix}. \tag{B.9}$$

This forms the basis for proving Theorem 3.

Step 2: analyzing the dynamics of the linear system (B.8). Before proceeding, we note that a linear system similar to (B.8) has been analyzed in Cen et al. [2022b, Section 4.2.2]. We intend to apply the following properties that have been derived therein:

$$x_{k+1} \leq A^{k+1}\left[x_0 + \gamma(\alpha^{-1}A - I)^{-1}y\right], \tag{B.10a}$$

$$\gamma(\alpha^{-1}A - I)^{-1}y = \begin{bmatrix} 0 \\ \|Q_\tau^{(0)} - \tau\xi^{(0)}\|_\infty \end{bmatrix}, \tag{B.10b}$$

$$A^{k+1} = \big((1-\alpha)\gamma + \alpha\big)^k \begin{bmatrix}\gamma \\ 1\end{bmatrix}\begin{bmatrix}1-\alpha & \alpha\end{bmatrix}. \tag{B.10c}$$

Substituting (B.10c) and (B.10b) into (B.10a) and rearranging terms, we reach

$$\begin{aligned} x_{k+1} &\leq \big((1-\alpha)\gamma + \alpha\big)^k \Big((1-\alpha)\big\|Q_\tau^\star - Q_\tau^{(0)}\big\|_\infty + \alpha\big\|Q_\tau^\star - \tau\xi^{(0)}\big\|_\infty + \alpha\big\|Q_\tau^{(0)} - \tau\xi^{(0)}\big\|_\infty\Big)\begin{bmatrix}\gamma \\ 1\end{bmatrix} \\ &\leq \big((1-\alpha)\gamma + \alpha\big)^k \Big(\big\|Q_\tau^\star - Q_\tau^{(0)}\big\|_\infty + 2\alpha\big\|Q_\tau^\star - \tau\xi^{(0)}\big\|_\infty\Big)\begin{bmatrix}\gamma \\ 1\end{bmatrix}, \end{aligned} \tag{B.11}$$

which taken together with the definition of x_{k+1} gives

$$\big\|Q_\tau^\star - Q_\tau^{(k+1)}\big\|_\infty \leq \gamma\big((1-\alpha)\gamma + \alpha\big)^k \Big(\big\|Q_\tau^\star - Q_\tau^{(0)}\big\|_\infty + 2\alpha\big\|Q_\tau^\star - \tau\xi^{(0)}\big\|_\infty\Big), \tag{B.12a}$$

$$\big\|Q_\tau^\star - \tau\xi^{(k+1)}\big\|_\infty \leq \big((1-\alpha)\gamma + \alpha\big)^k \Big(\big\|Q_\tau^\star - Q_\tau^{(0)}\big\|_\infty + 2\alpha\big\|Q_\tau^\star - \tau\xi^{(0)}\big\|_\infty\Big). \tag{B.12b}$$

Step 3: controlling $\big\|\pi_\tau^\star(s) - \pi^{(k+1)}(s)\big\|_1$ and $\big\|V_\tau^\star - V_\tau^{(k+1)}\big\|_\infty$. It remains to convert this result to an upper bound on $\big\|\pi_\tau^\star(s) - \pi^{(k+1)}(s)\big\|_1$ and $\big\|V_\tau^\star - V_\tau^{(k+1)}\big\|_\infty$. By virtue of Lemma 1, there exist two vectors $g_\tau^\star(s) \in \partial h_s\big(\pi_\tau^\star(s)\big)$, $g^{(k+1)}(s) \in \partial h_s\big(\pi^{(k+1)}(s)\big)$ and two scalars $c_s^\star, c_s^{(k+1)} \in \mathbb{R}$ that satisfy

$$\begin{cases} \tau^{-1}Q_\tau^\star(s) - c_s^\star 1 & = g_\tau^\star(s) \\ \xi^{(k+1)}(s,\cdot) - c_s^{(k+1)}1 & = g^{(k+1)}(s) \end{cases}.$$

It holds for all $s \in \mathcal{S}$ that

$$\begin{aligned} &V_\tau^\star(s) - V_\tau^{(k+1)}(s) \\ &= \big\langle Q_\tau^\star(s), \pi_\tau^\star(s)\big\rangle - \tau h_s(\pi_\tau^\star(s)) - \big\langle Q_\tau^{(k+1)}(s), \pi_\tau^{(k+1)}(s)\big\rangle + \tau h_s(\pi_\tau^{(k+1)}(s)) \\ &= \big\langle Q_\tau^\star(s) - Q_\tau^{(k+1)}(s), \pi_\tau^{(k+1)}(s)\big\rangle + \Big[\tau(h_s(\pi_\tau^{(k+1)}(s)) - h_s(\pi_\tau^\star(s))) - \big\langle Q_\tau^\star(s), \pi_\tau^{(k+1)}(s) - \pi_\tau^\star(s)\big\rangle\Big] \\ &\overset{\text{(i)}}{\leq} \big\langle Q_\tau^\star(s) - Q_\tau^{(k+1)}(s), \pi_\tau^{(k+1)}(s)\big\rangle + \big\langle \tau g^{(k+1)}(s) - Q_\tau^\star(s), \pi_\tau^{(k+1)}(s) - \pi_\tau^\star(s))\big\rangle \\ &= \big\langle Q_\tau^\star(s) - Q_\tau^{(k+1)}(s), \pi_\tau^{(k+1)}(s)\big\rangle + \big\langle \tau\xi^{(k+1)}(s) - Q_\tau^\star(s), \pi_\tau^{(k+1)}(s) - \pi_\tau^\star(s))\big\rangle \\ &\leq \big\|Q_\tau^\star(s) - Q_\tau^{(k+1)}(s)\big\|_\infty + 2\big\|Q_\tau^\star(s) - \tau\xi^{(k+1)}(s)\big\|_\infty, \end{aligned} \tag{B.13}$$

where (i) results from $h_s(\pi_\tau^{(k+1)}(s)) - h_s(\pi_\tau^\star(s)) \leq \big\langle g^{(k+1)}(s), \pi_\tau^{(k+1)}(s) - \pi_\tau^\star(s)\big\rangle$. Plugging (B.12) into (B.13) completes the proof for (3.15b).

When h_s is 1-strongly convex w.r.t. the ℓ_1 norm, we can invoke the strong monotonicity property of a strongly convex function [Beck, 2017, Theorem 5.24] to obtain

$$\begin{aligned}\big\|\pi_\tau^\star(s)-\pi^{(k+1)}(s)\big\|_1^2 &\le \big\langle \pi_\tau^\star(s)-\pi^{(k+1)}(s), g_\tau^\star(s)-g^{(k+1)}(s)\big\rangle \\ &= \big\langle \pi_\tau^\star(s)-\pi^{(k+1)}(s), g_\tau^\star(s)+c_s^\star 1-g^{(k+1)}(s)-c_s^{(k+1)}1\big\rangle \\ &\le \big\|\pi_\tau^\star(s)-\pi^{(k+1)}(s)\big\|_1\big\|g_\tau^\star(s)+c_s^\star 1-g^{(k+1)}(s)-c_s^{(k+1)}1\big\|_\infty \\ &= \tau^{-1}\big\|\pi_\tau^\star(s)-\pi^{(k+1)}(s)\big\|_1\big\|Q_\tau^\star(s)-\tau\xi^{(k+1)}(s)\big\|_\infty, \end{aligned} \tag{B.14}$$

where the second line is valid since $\langle \pi_\tau^\star(s),1\rangle=\langle \pi^{(k+1)}(s),1\rangle=1$. This taken together with (B.12) gives rise to the advertised bound

$$\begin{aligned}\big\|\pi_\tau^\star(s)-\pi^{(k+1)}(s)\big\|_1 &\le \tau^{-1}\big\|Q_\tau^\star(s)-\tau\xi^{(k+1)}(s)\big\|_\infty \\ &\le \tau^{-1}\big((1-\alpha)\gamma+\alpha\big)^k\left(\big\|Q_\tau^\star-Q_\tau^{(0)}\big\|_\infty+2\alpha\big\|Q_\tau^\star-\tau\xi^{(0)}\big\|_\infty\right).\end{aligned}$$

B.2 Proof of key lemmas

In this section, we collect the proof of several key lemmas. Here and throughout, we use $\mathbb{E}_\pi[\cdot]$ to denote the expectation over the randomness of the MDP induced by policy π. We shall follow the notation convention in (B.1) throughout. In addition, to further simplify notation, we shall abuse the notation by letting

$$D_{h_s}(\widetilde{\pi},\pi;\xi) \coloneqq D_{h_s}\big(\widetilde{\pi}(\cdot\,|\,s),\pi(\cdot\,|\,s);\xi(s,\cdot)\big) \tag{B.15a}$$
$$D_{h_s}(p,\pi;\xi) \coloneqq D_{h_s}\big(p,\pi(\cdot\,|\,s);\xi(s,\cdot)\big) \tag{B.15b}$$
$$D_{h_s}(\pi,p;\xi) \coloneqq D_{h_s}\big(\pi(\cdot\,|\,s),p;\xi(s,\cdot)\big) \tag{B.15c}$$

for any policy π and $\widetilde{\pi}$ and any $p\in\Delta(\mathcal{A})$, whenever it is clear from the context.

B.2.1 Proof of Lemma 1

We start by relaxing the probability simplex constraint (i.e., $p\in\Delta(\mathcal{A})$) in (3.14a) with a simpler linear constraint $\sum_{a\in\mathcal{A}}p(a)=1$ as follows

$$\begin{array}{ll}\text{minimize}_{p\in\mathbb{R}^{|\mathcal{A}|}} & -\eta\big\langle Q_\tau^{(k)}(s),p\big\rangle+\eta\tau h_s(p)+D_{h_s}\big(p,\pi^{(k)};\xi^{(k)}\big)\\ \text{subject to} & \sum_{a\in\mathcal{A}}p(a)=1.\end{array} \tag{B.16}$$

To justify the validity of dropping the non-negative constraint, we note that for any p obeying $p(a)<0$ for some $a\in\mathcal{A}$, our assumption on h_s (see Assumption 1) leads to $h_s(p)=\infty$, which cannot possibly be the optimal solution. This confirms the equivalence between (3.14a) and (B.16).

Observe that the Lagrangian w.r.t. (B.16) is given by

$$\begin{aligned}&\mathcal{L}_s\big(p,\lambda_s^{(k)}\big)\\ &=-\eta\big\langle Q_\tau^{(k)}(s),p\big\rangle+\eta\tau h_s(p)+h_s(p)-h_s\big(\pi^{(k)}(s)\big)-\big\langle p-\pi^{(k)}(s),\xi^{(k)}(s)\big\rangle+\lambda_s^{(k)}\left(\sum_{a\in\mathcal{A}}p(a)-1\right),\end{aligned}$$

where $\lambda_s^{(k)}\in\mathbb{R}$ denotes the Lagrange multiplier associated with the constraint $\sum_{a\in\mathcal{A}}p(a)=1$. Given that $\pi^{(k+1)}(s)$ is the solution to (3.14a) and hence (B.16), the optimality condition requires that

$$0\in\partial_p\mathcal{L}_s\big(p,\lambda_s^{(k)}\big)\Big|_{p=\pi^{(k+1)}(s)}=-\eta Q_\tau^{(k)}(s)+(1+\eta\tau)\partial h_s\big(\pi^{(k+1)}(s)\big)-\xi^{(k)}(s)+\lambda_s^{(k)}1.$$

Rearranging terms and making use of the construction (3.11), we are left with

$$\xi^{(k+1)}(s) - \frac{\lambda_s^{(k)}}{1+\eta\tau}1 = \frac{1}{1+\eta\tau}\left[\eta Q_\tau^{(k)}(s) + \xi^{(k)}(s) - \lambda_s^{(k)}1\right] \in \partial h_s\big(\pi^{(k+1)}(s)\big),$$

thus concluding the proof of the first claim (3.12).

We now turn to the second claim (3.13). In view of the property (B.5), we have

$$\pi_\tau^\star(s) = \arg\min_{p\in\Delta(\mathcal{A})} -\big\langle Q_\tau^\star(s), p\big\rangle + \tau h_s(p).$$

This optimization problem is equivalent to

$$\begin{array}{ll} \text{minimize}_{p\in\mathbb{R}^{|\mathcal{A}|}} & -\big\langle Q_\tau^\star(s), p\big\rangle + \tau h_s(p), \\ \text{subject to} & \sum_{a\in\mathcal{A}} p(a) = 1, \end{array} \tag{B.17}$$

which can be verified by repeating a similar argument for (B.16). The Lagrangian associated with (B.17) is

$$\mathcal{L}_s\big(p, \lambda_s^\star\big) = -\big\langle Q_\tau^\star(s), p\big\rangle + \tau h_s(p) + \lambda_s^\star\left(\sum_{a\in\mathcal{A}} p(a) - 1\right),$$

where $\lambda_s^\star \in \mathbb{R}$ denotes the Lagrange multiplier. Therefore, the first-order optimality condition requires that

$$0 \in \partial_p \mathcal{L}_s\big(p, \lambda_s^\star\big)\Big|_{p=\pi_\tau^\star(s)} = -Q_\tau^\star(s) + \tau\partial h_s\big(\pi_\tau^\star(s)\big) + \lambda_s^\star 1,$$

which immediately finishes the proof.

B.2.2 Proof of Lemma 8

We start by introducing the performance difference lemma that has previously been derived in Lan [2023, Lemma 2]. For the sake of self-containedness, we include a proof of this lemma in Appendix B.2.2.

Lemma 11 (Performance difference)**.** *For any two policies π and π', we have*

$$V_\tau^{\pi'}(s) - V_\tau^\pi(s) = \frac{1}{1-\gamma}\mathop{\mathbb{E}}_{s'\sim d_s^{\pi'}}\Big[\big\langle Q_\tau^\pi(s'), \pi'(s') - \pi(s')\big\rangle - \tau h_{s'}\big(\pi'(s')\big) + \tau h_{s'}\big(\pi(s')\big)\Big], \tag{B.18}$$

where d_s^π has been defined in (2.4).

Armed with Lemma 11, one can readily rewrite the difference $V_\tau^{(k+1)}(s) - V_\tau^{(k)}(s)$ between two consecutive iterates as follows

$$\begin{aligned} &V_\tau^{(k+1)}(s) - V_\tau^{(k)}(s) \\ &\quad = \frac{1}{1-\gamma}\mathop{\mathbb{E}}_{s'\sim d_s^{(k+1)}}\Big[\big\langle Q_\tau^{(k)}(s'), \pi^{(k+1)}(s') - \pi^{(k)}(s')\big\rangle - \tau h_{s'}\big(\pi^{(k+1)}(s')\big) + \tau h_{s'}\big(\pi^{(k)}(s')\big)\Big]. \end{aligned} \tag{B.19}$$

It then comes down to studying the right-hand side of the relation (B.19), which can be accomplished via the following "three-point" lemma. The proof of this lemma can be found in Appendix B.2.2.

Lemma 12. *For any $s \in \mathcal{S}$ and any vector $p \in \Delta(\mathcal{A})$, we have*

$$\begin{aligned}&(1+\eta\tau)D_{h_s}\big(p, \pi^{(k+1)}; \xi^{(k+1)}\big) + D_{h_s}\big(\pi^{(k+1)}, \pi^{(k)}; \xi^{(k)}\big) - D_{h_s}\big(p, \pi^{(k)}; \xi^{(k)}\big) \\ &\quad = \eta\left[\big\langle Q_\tau^{(k)}(s), \pi^{(k+1)}(s) - p\big\rangle + \tau h_s(p) - \tau h_s\big(\pi^{(k+1)}(s)\big)\right].\end{aligned}$$

Taking $p = \pi^{(k)}(s)$ in Lemma 12 and combining it with (B.19), we arrive at

$$\begin{aligned}&V_\tau^{(k+1)}(s) - V_\tau^{(k)}(s) \\ &= \frac{1}{(1-\gamma)\eta} \mathop{\mathbb{E}}_{s'\sim d_s^{(k+1)}}\left[(1+\eta\tau)D_{h_{s'}}\big(\pi^{(k)}, \pi^{(k+1)}; \xi^{(k+1)}\big) + D_{h_{s'}}\big(\pi^{(k+1)}, \pi^{(k)}; \xi^{(k)}\big)\right] \geq 0\end{aligned}$$

for any $s \in \mathcal{S}$, thus establishing the advertised pointwise monotonicity w.r.t. the regularized value function.

When it comes to the regularized Q-function, it is readily seen from the definition (3.3a) that

$$\begin{aligned}Q_\tau^{(k+1)}(s,a) &= r(s,a) + \gamma \mathop{\mathbb{E}}_{s'\sim P(\cdot|s,a)}\left[V_\tau^{(k+1)}(s')\right] \\ &\geq r(s,a) + \gamma \mathop{\mathbb{E}}_{s'\sim P(\cdot|s,a)}\left[V_\tau^{(k)}(s')\right] = Q_\tau^{(k)}(s,a)\end{aligned}$$

for any $(s,a) \in \mathcal{S}\times\mathcal{A}$, where the last line is valid since $V_\tau^{(k+1)} \geq V_\tau^{(k)}$. This concludes the proof.

Proof of Lemma 11

For any two policies π' and π, it follows from the definition (3.1) of $V_\tau^\pi(s)$ that

$$\begin{aligned}V_\tau^{\pi'}(s) - V_\tau^\pi(s) &= \mathbb{E}_{\pi'}\left[\sum_{t=0}^\infty \gamma^t\Big[r(s_t,a_t) - \tau h_{s_t}\big(\pi'(s_t)\big)\Big] \,\Big|\, s_0 = s\right] - V_\tau^\pi(s) \\ &= \mathbb{E}_{\pi'}\left[\sum_{t=0}^\infty \gamma^t\Big[r(s_t,a_t) - \tau h_{s_t}\big(\pi'(s_t)\big) + V_\tau^\pi(s_t) - V_\tau^\pi(s_t)\Big] \,\Big|\, s_0 = s\right] - V_\tau^\pi(s) \\ &= \mathbb{E}_{\pi'}\left[\sum_{t=0}^\infty \gamma^t\Big[r(s_t,a_t) - \tau h_{s_t}\big(\pi'(s_t)\big) + \gamma V_\tau^\pi(s_{t+1}) - V_\tau^\pi(s_t)\Big] \,\Big|\, s_0 = s\right] + \mathbb{E}_{\pi'}\left[V_\tau^\pi(s_0) \,\Big|\, s_0 = s\right] - V_\tau^\pi(s) \\ &= \mathbb{E}_{\pi'}\left[\sum_{t=0}^\infty \gamma^t\Big[r(s_t,a_t) - \tau h_{s_t}\big(\pi'(s_t)\big) + \gamma V_\tau^\pi(s_{t+1}) - V_\tau^\pi(s_t)\Big] \,\Big|\, s_0 = s\right] \\ &= \mathbb{E}_{\pi'}\left[\sum_{t=0}^\infty \gamma^t\Big[r(s_t,a_t) - \tau h_{s_t}\big(\pi(s_t)\big) + \gamma V_\tau^\pi(s_{t+1}) - V_\tau^\pi(s_t) - \tau h_{s_t}\big(\pi'(s_t)\big) + \tau h_{s_t}\big(\pi(s_t)\big)\Big] \,\Big|\, s_0 = s\right] \\ &= \mathbb{E}_{\pi'}\left[\sum_{t=0}^\infty \gamma^t\Big[Q_\tau^\pi(s_t,a_t) - \tau h_{s_t}\big(\pi(s_t)\big) - V_\tau^\pi(s_t) - \tau h_{s_t}\big(\pi'(s_t)\big) + \tau h_{s_t}\big(\pi(s_t)\big)\Big] \,\Big|\, s_0 = s\right] \\ &= \frac{1}{1-\gamma}\mathop{\mathbb{E}}_{s'\sim d_s^{\pi'}}\Big[\big\langle Q_\tau^\pi(s'), \pi'(s') - \pi(s')\big\rangle - \tau h_{s'}\big(\pi'(s')\big) + \tau h_{s'}\big(\pi(s')\big)\Big], \qquad \text{(B.20)}\end{aligned}$$

where the penultimate line comes from the definition (3.3a). To see why the last line of (B.20) is valid, we make note of the following identity

$$\begin{aligned}
&\mathop{\mathbb{E}}_{a_t\sim\pi'(s_t)}\Big[Q_\tau^\pi(s_t,a_t)-\tau h_{s_t}\big(\pi(s_t)\big)-V_\tau^\pi(s_t)\Big] \\
&\quad=\mathop{\mathbb{E}}_{a_t\sim\pi'(s_t)}\Big[Q_\tau^\pi(s_t,a_t)-\tau h_{s_t}\big(\pi(s_t)\big)\Big]-\mathop{\mathbb{E}}_{a_t\sim\pi(s_t)}\Big[Q_\tau^\pi(s_t,a_t)-\tau h_{s_t}\big(\pi(s_t)\big)\Big] \\
&\quad=\Big\langle Q_\tau^\pi(s_t)-\tau h_{s_t}\big(\pi(s_t)\big)\cdot 1,\pi'(s_t)-\pi(s_t)\Big\rangle \\
&\quad=\Big\langle Q_\tau^\pi(s_t),\pi'(s_t)-\pi(s_t)\Big\rangle, \qquad\qquad (B.21)
\end{aligned}$$

where the first identity results from the relation (3.3b), and the last relation holds since $1^\top\pi'(s_t)=1^\top\pi(s_t)=1$. The last line of (B.20) then follows immediately from the relation (B.21) and the definition (2.4) of d_s^π.

Proof of Lemma 12

For any state $s\in\mathcal{S}$, we make the observation that

$$\begin{aligned}
&D_{h_s}\big(p,\pi^{(k)};\xi^{(k)}\big)=h_s(p)-h_s\big(\pi^{(k)}(s)\big)-\big\langle p-\pi^{(k)}(s),\xi^{(k)}(s)\big\rangle \\
&\quad=h_s(p)-h_s\big(\pi^{(k+1)}(s)\big)-\big\langle p-\pi^{(k+1)}(s),\xi^{(k)}(s)\big\rangle \\
&\qquad+h_s\big(\pi^{(k+1)}(s)\big)-h_s\big(\pi^{(k)}(s)\big)-\big\langle\pi^{(k+1)}(s)-\pi^{(k)}(s),\xi^{(k)}(s)\big\rangle \\
&\quad=h_s(p)-h_s\big(\pi^{(k+1)}(s)\big)-\big\langle p-\pi^{(k+1)}(s),\xi^{(k+1)}(s)\big\rangle \\
&\qquad+h_s\big(\pi^{(k+1)}(s)\big)-h_s\big(\pi^{(k)}(s)\big)-\big\langle\pi^{(k+1)}(s)-\pi^{(k)}(s),\xi^{(k)}(s)\big\rangle \\
&\qquad+\big\langle p-\pi^{(k+1)}(s),\xi^{(k+1)}(s)-\xi^{(k)}(s)\big\rangle\big\rangle \\
&\quad=D_{h_s}\big(p,\pi^{(k+1)};\xi^{(k+1)}\big)+D_{h_s}\big(\pi^{(k+1)},\pi^{(k)};\xi^{(k)}\big)+\big\langle p-\pi^{(k+1)}(s),\xi^{(k+1)}(s)-\xi^{(k)}(s)\big\rangle \\
&\quad=D_{h_s}\big(p,\pi^{(k+1)};\xi^{(k+1)}\big)+D_{h_s}\big(\pi^{(k+1)},\pi^{(k)};\xi^{(k)}\big)+\big\langle p-\pi^{(k+1)}(s),\eta Q_\tau^{(k)}(s)-\eta\tau\xi^{(k+1)}(s)\big\rangle,
\end{aligned}$$

where the first and the fourth steps invoke the definition (3.9) of the generalized Bregman divergence and the last line results from the update rule (3.14c). Rearranging terms, we are left with

$$\begin{aligned}
&\eta\big\langle Q_\tau^{(k)}(s),\pi^{(k+1)}(s)-p\big\rangle \\
&\qquad=\Big\{D_{h_s}\big(p,\pi^{(k+1)};\xi^{(k+1)}\big)+D_{h_s}\big(\pi^{(k+1)},\pi^{(k)};\xi^{(k)}\big)-D_{h_s}\big(p,\pi^{(k)};\xi^{(k)}\big)\Big\} \\
&\qquad\quad+\eta\tau\big\langle\xi^{(k+1)}(s),\pi^{(k+1)}(s)-p\big\rangle.
\end{aligned}$$

Adding the term $\eta\tau\big\{h_s(p)-h_s\big(\pi^{(k+1)}(s)\big)\big\}$ to both sides of this identity leads to

$$\begin{aligned}
&\eta\Big[\big\langle Q_\tau^{(k)}(s),\pi^{(k+1)}(s)-p\big\rangle+\tau h_s(p)-\tau h_s\big(\pi^{(k+1)}(s)\big)\Big] \\
&\qquad=\Big\{D_{h_s}\big(p,\pi^{(k+1)};\xi^{(k+1)}\big)+D_{h_s}\big(\pi^{(k+1)},\pi^{(k)};\xi^{(k)}\big)-D_{h_s}\big(p,\pi^{(k)};\xi^{(k)}\big)\Big\} \\
&\qquad\quad+\eta\tau\Big(h_s(p)-h_s\big(\pi^{(k+1)}(s)\big)-\big\langle\xi^{(k+1)}(s),p-\pi^{(k+1)}(s)\big\rangle\Big) \\
&\qquad=(1+\eta\tau)D_{h_s}\big(p,\pi^{(k+1)};\xi^{(k+1)}\big)+D_{h_s}\big(\pi^{(k+1)},\pi^{(k)};\xi^{(k)}\big)-D_{h_s}\big(p,\pi^{(k)};\xi^{(k)}\big)
\end{aligned}$$

as claimed, where the last line makes use of the definition (3.9).

B.2.3 Proof of Lemma 9

In the sequel, we shall prove each claim in Lemma 9 separately.

Proof of the contraction property (B.4). For any $Q_1, Q_2 \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$, the definition (B.3) of the generalized Bellman operator obeys

$$\begin{aligned}
\mathcal{T}_{\tau,h}(Q_1) - \mathcal{T}_{\tau,h}(Q_2) &= \gamma \mathop{\mathbb{E}}_{s'\sim P(\cdot|s,a)} \left[\max_{p\in\Delta(\mathcal{A})} \Big\{ \langle Q_1(s'), p\rangle - \tau h_{s'}(p) \Big\} - \max_{p\in\Delta(\mathcal{A})} \Big\{ \langle Q_2(s'), p\rangle - \tau h_{s'}(p) \Big\} \right] \\
&\overset{\text{(a)}}{\le} \gamma \mathop{\mathbb{E}}_{s'\sim P(\cdot|s,a)} \left[\max_{p\in\Delta(\mathcal{A})} \big\langle Q_1(s') - Q_2(s'), p \big\rangle \right] \\
&\le \gamma \mathop{\mathbb{E}}_{s'\sim P(\cdot|s,a)} \left[\max_{p:\|p\|_1=1} \|Q_1 - Q_2\|_\infty \|p\|_1 \right] \\
&= \gamma \|Q_1 - Q_2\|_\infty,
\end{aligned}$$

where (a) arises from the elementary fact $\max_x f(x) - \max_x g(x) \le \max_x \big(f(x) - g(x)\big)$.

Proof of the fixed point property (B.5). Towards this, let us first define

$$\pi^\dagger(s) \coloneqq \arg \max_{p_s\in\Delta(\mathcal{A})} \mathop{\mathbb{E}}_{a\sim p_s} \Big[Q^\star_\tau(s,a) - \tau h_s\big(p(s)\big) \Big]. \tag{B.22}$$

Then it can be easily verified that

$$\begin{aligned}
Q^\star_\tau(s,a) &= r(s,a) + \gamma \mathop{\mathbb{E}}_{s_1\sim P(\cdot|s,a)} \left[\mathop{\mathbb{E}}_{a_1\sim\pi^\star(s_1)} \Big[Q^\star_\tau(s_1,a_1) - \tau h_{s_1}\big(\pi^\star(s_1)\big) \Big] \right] \\
&\le r(s,a) + \gamma \mathop{\mathbb{E}}_{s_1\sim P(\cdot|s,a)} \left[\mathop{\mathbb{E}}_{a_1\sim\pi^\dagger(s_1)} \Big[Q^\star_\tau(s_1,a_1) - \tau h_{s_1}\big(\pi^\dagger(s_1)\big) \Big] \right],
\end{aligned} \tag{B.23}$$

where the first identity results from (3.3), and the second line arises from the maximizing property of $\pi^\dagger$ (see (B.22)).

Note that the right-hand side of (B.23) involves the term $Q^\star_\tau(s_1,a_1)$, which can be further upper bounded via the same argument for (B.23). Successively repeating this upper bound argument (and the expansion) eventually allows one to obtain

$$Q^\star_\tau(s,a) \le r(s,a) + \gamma \mathbb{E}_{\pi^\dagger} \left[\sum_{t=1}^{\infty} \gamma^{t-1} \Big\{ r(s_t,a_t) - \tau h_{s_t}\big(\pi^\dagger(s_t)\big) \Big\} \,\Big|\, s_0 = s, a_0 = a \right] = Q^{\pi^\dagger}_\tau(s,a).$$

However, the fact that $\pi^\star$ is the optimal policy necessarily implies the following reverse inequality:

$$Q^\star_\tau(s,a) \ge Q^{\pi^\dagger}_\tau(s,a).$$

Therefore, one must have

$$Q^\star_\tau(s,a) = Q^{\pi^\dagger}_\tau(s,a). \tag{B.24}$$

To finish up, it suffices to show that $Q_\tau^{\pi^\dagger} = \mathcal{T}_{\tau,h}(Q_\tau^\star)$. To this end, it is observed that

$$\begin{aligned}
Q_\tau^{\pi^\dagger}(s,a) &= r(s,a) + \gamma \mathop{\mathbb{E}}_{s_1\sim P(\cdot|s,a)}\left[\mathop{\mathbb{E}}_{a_1\sim\pi^\dagger(s_1)}\left[Q_\tau^{\pi^\dagger}(s_1,a_1) - \tau h_{s_1}\big(\pi^\dagger(s_1)\big)\right]\right] \\
&\overset{\text{(b)}}{=} r(s,a) + \gamma \mathop{\mathbb{E}}_{s_1\sim P(\cdot|s,a)}\left[\mathop{\mathbb{E}}_{a_1\sim\pi^\dagger(s_1)}\left[Q_\tau^{\star}(s_1,a_1) - \tau h_{s_1}\big(\pi^\dagger(s_1)\big)\right]\right] \\
&\overset{\text{(c)}}{=} r(s,a) + \gamma \mathop{\mathbb{E}}_{s_1\sim P(\cdot|s,a)}\left[\max_{p\in\Delta(\mathcal{A})}\big\langle Q_\tau^\star(s_1,a_1), p\big\rangle - \tau h_{s_1}(p)\right] \\
&= \mathcal{T}_{\tau,h}(Q_\tau^\star)(s,a),
\end{aligned}$$

where (b) utilizes the fact (B.24), (c) follows from the definition (B.22) of $\pi^\dagger$, and the last identity is a consequence of the definition (B.3) of $\mathcal{T}_{\tau,h}$. The above results taken collectively demonstrate that $Q_\tau^\star = \mathcal{T}_{\tau,h}(Q_\tau^\star)$ as claimed.

B.2.4 Proof of Lemma 10

Recall that $Q_\tau^{(k+1)} = Q_\tau^{\pi^{(k+1)}}$. In view of the relation (3.3), one obtains

$$\begin{aligned}
Q_\tau^{(k+1)}(s,a) &= r(s,a) + \gamma \mathop{\mathbb{E}}_{s'\sim P(\cdot|s,a)}\left[V_\tau^{(k+1)}(s')\right] \\
&= r(s,a) + \gamma \mathop{\mathbb{E}}_{s'\sim P(\cdot|s,a)}\left[\mathop{\mathbb{E}}_{a'\sim\pi^{(k+1)}(s')}\left[Q_\tau^{(k+1)}(s',a') - \tau h_{s'}\big(\pi^{(k+1)}(s')\big)\right]\right] \\
&= r(s,a) + \gamma \mathop{\mathbb{E}}_{s'\sim P(\cdot|s,a)}\left[\big\langle Q_\tau^{(k+1)}(s'), \pi^{(k+1)}(s')\big\rangle - \tau h_{s'}\big(\pi^{(k+1)}(s')\big)\right].
\end{aligned}$$

This combined with the fixed-point condition (B.5) allows us to derive

$$\begin{aligned}
&Q_\tau^\star(s,a) - Q_\tau^{(k+1)}(s,a) \\
&= \mathcal{T}_{\tau,h}(Q_\tau^\star)(s,a) - \left\{r(s,a) + \gamma \mathop{\mathbb{E}}_{s'\sim P(\cdot|s,a)}\left[\big\langle Q_\tau^{(k+1)}(s'), \pi^{(k+1)}(s')\big\rangle - \tau h_{s'}\big(\pi^{(k+1)}(s')\big)\right]\right\} \\
&= \mathcal{T}_{\tau,h}(Q_\tau^\star)(s,a) - \left\{r(s,a) + \gamma \mathop{\mathbb{E}}_{s'\sim P(\cdot|s,a)}\left[\big\langle \tau\xi^{(k+1)}(s'), \pi^{(k+1)}(s')\big\rangle - \tau h_{s'}\big(\pi^{(k+1)}(s')\big)\right]\right\} \\
&\qquad - \gamma \mathop{\mathbb{E}}_{s'\sim P(\cdot|s,a), a'\sim\pi^{(k+1)}(s')}\left[Q_\tau^{(k+1)}(s',a') - \tau\xi^{(k+1)}(s',a')\right].
\end{aligned} \tag{B.25}$$

In what follows, we control each term on the right-hand side of (B.25) separately.

Step 1: bounding the 1st term on the right-hand side of (B.25). Lemma 1 tells us that

$$\xi^{(k+1)}(s) - c_s^{(k+1)}1 \in \partial h_s(\pi^{(k+1)}(s))$$

for some scalar $c_s^{(k+1)} \in \mathbb{R}$. This important property allows one to derive

$$0 \in -\xi^{(k+1)}(s) + c_s^{(k+1)}1 + \partial h_s\big(\pi^{(k+1)}(s)\big) = \partial\mathcal{L}_{k+1,s}\big(\pi^{(k+1)}(s); c_s^{(k+1)}\big) \tag{B.26}$$

where

$$\mathcal{L}_{k+1,s}(p;\lambda) \coloneqq \underbrace{-\big\langle\xi^{(k+1)}(s), p\big\rangle + h_s\big(p\big)}_{=:\, f_{k+1,s}(p)} + \lambda\, 1^\top p.$$

Recognizing that the function $f_{k+1,s}(\cdot)$ is convex in p, we can view $\mathcal{L}_{k+1,s}(p;\lambda)$ as the Lagrangian of the following constrained convex problem with Lagrangian multiplier $\lambda \in \mathbb{R}$:

$$\underset{p:1^\top p=1}{\text{minimize}} \quad f_{k+1,s}(p) = -\big\langle \xi^{(k+1)}(s), p\big\rangle + h_s\big(p\big). \tag{B.27}$$

The condition (B.26) can then be interpreted as the optimality condition w.r.t. the program (B.27) and $\pi^{(k+1)}(s)$, meaning that

$$f_{k+1,s}\big(\pi^{(k+1)}(s)\big) = \min_{p:1^\top p=1} f_{k+1,s}(p),$$

or equivalently,

$$\big\langle \xi^{(k+1)}(s), \pi^{(k+1)}(s)\big\rangle - h_s\big(\pi^{(k+1)}(s)\big) = \max_{p:1^\top p=1} \big\langle \xi^{(k+1)}(s), p\big\rangle - h_s(p). \tag{B.28}$$

In addition, for any vector p that does not obey $p \geq 0$, Assumption 1 implies that $h_s(p) = \infty$, and hence p cannot possibly be the optimal solution to $\max_{p\in\Delta(\mathcal{A})} \big\langle \xi^{(k+1)}(s), p\big\rangle - h_s(p)$. This together with (B.28) essentially implies that

$$\big\langle \xi^{(k+1)}(s), \pi^{(k+1)}(s)\big\rangle - h_s\big(\pi^{(k+1)}(s)\big) = \max_{p\in\Delta(\mathcal{A})} \big\langle \xi^{(k+1)}(s), p\big\rangle - h_s(p). \tag{B.29}$$

As a consequence, we arrive at

$$\begin{aligned}
&\mathcal{T}_{\tau,h}(Q_\tau^\star)(s,a) - \left\{ r(s,a) + \gamma \mathop{\mathbb{E}}_{s'\sim P(\cdot|s,a)} \Big[\big\langle \tau\xi^{(k+1)}(s'), \pi^{(k+1)}(s')\big\rangle - \tau h_{s'}\big(\pi^{(k+1)}(s')\big)\Big]\right\} \\
&\quad = \mathcal{T}_{\tau,h}(Q_\tau^\star)(s,a) - \left\{ r(s,a) + \gamma \mathop{\mathbb{E}}_{s'\sim P(\cdot|s,a)} \Big[\max_{p\in\Delta(\mathcal{A})} \Big\{ \big\langle \tau\xi^{(k+1)}(s'), p\big\rangle - \tau h_{s'}(p)\Big\}\Big]\right\} \\
&\quad = \mathcal{T}_{\tau,h}(Q_\tau^\star)(s,a) - \mathcal{T}_{\tau,h}(\tau\xi^{(k+1)})(s,a) \\
&\quad \leq \gamma \big\| Q_\tau^\star - \tau\xi^{(k+1)}\big\|_\infty,
\end{aligned} \tag{B.30}$$

where the last step results from the contraction property (B.4) in Lemma 9.

Step 2: bounding the 2nd term on the right-hand side of (B.25). Recall that $\alpha = \frac{1}{1+\eta\tau}$. Invoking the monotonicity property in Lemma 8 and the update rule (3.14c), we obtain

$$\begin{aligned}
Q_\tau^{(k+1)}(s,a) - \tau\xi^{(k+1)}(s,a) &= \alpha\Big\{ Q_\tau^{(k+1)}(s,a) - \tau\xi^{(k)}(s,a)\Big\} + (1-\alpha)\Big\{ Q_\tau^{(k+1)}(s,a) - Q_\tau^{(k)}(s,a)\Big\} \\
&\geq \alpha\Big\{ Q_\tau^{(k)}(s,a) - \tau\xi^{(k)}(s,a)\Big\}.
\end{aligned}$$

Repeating this lower bound argument then yields

$$\begin{aligned}
Q_\tau^{(k+1)}(s,a) - \tau\xi^{(k+1)}(s,a) &\geq \alpha^{k+1}\Big\{ Q_\tau^{(0)}(s,a) - \tau\xi^{(0)}(s,a)\Big\} \\
&\geq -\alpha^{k+1}\big\| Q_\tau^{(0)} - \tau\xi^{(0)}\big\|_\infty,
\end{aligned}$$

thus revealing that

$$-\mathop{\mathbb{E}}_{s'\sim P(\cdot|s,a), a'\sim\pi^{k+1}(s')} \Big[Q_\tau^{(k+1)}(s',a') - \tau\xi^{(k+1)}(s',a')\Big] \leq \alpha^{k+1}\big\| Q_\tau^{(0)} - \tau\xi^{(0)}\big\|_\infty. \tag{B.31}$$

Step 3: putting all this together. Substituting (B.30) and (B.31) into (B.25) gives

$$0 \leq Q_\tau^\star(s,a) - Q_\tau^{(k+1)}(s,a) \leq \gamma\big\| Q_\tau^\star - \tau\xi^{(k+1)}\big\|_\infty + \alpha^{k+1}\big\| Q_\tau^{(0)} - \tau\xi^{(0)}\big\|_\infty \tag{B.32}$$

for all $(s,a) \in \mathcal{S}\times\mathcal{A}$, thus concluding the proof.

Appendix C

Proofs for Chapter 4

C.1 Analysis for entropy-regularized matrix games

Before embarking on the main proof, it is useful to first consider the update rule (4.6) that underlies both PU and OMWU, which is reproduced below for convenience:

$$\begin{cases} \mu^{(t+1)}(a) \propto \mu^{(t)}(a)^{1-\eta\tau} \exp(\eta[Az_2]_a), & \text{for all } a \in \mathcal{A}, \\ \nu^{(t+1)}(b) \propto \nu^{(t)}(b)^{1-\eta\tau} \exp(-\eta[A^\top z_1]_b), & \text{for all } b \in \mathcal{B}, \end{cases} \tag{C.1}$$

where $z_1 \in \Delta(\mathcal{A})$ and $z_2 \in \Delta(\mathcal{B})$. These updates satisfy the following property, whose proof is provided in Appendix C.2.1.

Lemma 13. *Denote $\zeta^{(t)} = (\mu^{(t)}, \nu^{(t)})$ and $\zeta(z) = (z_1, z_2)$. The update rule (C.1) satisfies:*

$$\left\langle \log \mu^{(t+1)} - (1-\eta\tau)\log \mu^{(t)} - \eta\tau \log \mu_\tau^\star, z_1 - \mu_\tau^\star \right\rangle = \eta(\mu_\tau^\star - z_1)^\top A(\nu_\tau^\star - z_2), \tag{C.2a}$$

$$\left\langle \log \nu^{(t+1)} - (1-\eta\tau)\log \nu^{(t)} - \eta\tau \log \nu_\tau^\star, z_2 - \nu_\tau^\star \right\rangle = -\eta(\nu_\tau^\star - z_1)^\top A(\nu_\tau^\star - z_2), \tag{C.2b}$$

and

$$\left\langle \log \zeta^{(t+1)} - (1-\eta\tau)\log \zeta^{(t)} - \eta\tau \log \zeta_\tau^\star, \zeta(z) - \zeta_\tau^\star \right\rangle = 0. \tag{C.3}$$

As we shall see, the above lemma plays a crucial role in establishing the claimed convergence results. The next lemma gives some basic decompositions related to the game values that are helpful.

Lemma 14. *For every $(\mu, \nu) \in \Delta(\mathcal{A}) \times \Delta(\mathcal{B})$, the following relations hold*

$$f_\tau(\mu_\tau^\star, \nu) - f_\tau(\mu, \nu_\tau^\star) = \tau \mathsf{KL}\left(\zeta \,\|\, \zeta_\tau^\star\right), \tag{C.4a}$$

$$f_\tau(\mu, \nu) - f_\tau(\mu_\tau^\star, \nu_\tau^\star) = (\mu_\tau^\star - \mu)^\top A(\nu_\tau^\star - \nu) + \tau \mathsf{KL}\left(\nu \,\|\, \nu_\tau^\star\right) - \tau \mathsf{KL}\left(\mu \,\|\, \mu_\tau^\star\right). \tag{C.4b}$$

In addition, we also make record of the following elementary lemma that is used frequently.

Lemma 15 ([Mei et al., 2020b, Lemma 27])**.** *For any $\mu_1, \mu_2 \in \Delta(\mathcal{A})$ satisfying*

$$\mu_1(a) \propto \exp(x_1(a)) \quad \text{and} \quad \mu_2(a) \propto \exp(x_2(a))$$

for some $x_1, x_2 \in \mathbb{R}^{|\mathcal{A}|}$, we have

$$\mathsf{KL}\left(\mu_1 \,\|\, \mu_2\right) \le \frac{1}{2} \left\| x_1 - x_2 - c \cdot \mathbf{1} \right\|_\infty^2,$$

for all $c \in \mathbb{R}$.

C.1.1 Proof of Proposition 1

Setting $\zeta(z) = \zeta^{(t+1)}$ in Lemma 13, we have

$$\left\langle \log \zeta^{(t+1)} - (1-\eta\tau)\log \zeta^{(t)} - \eta\tau \log \zeta_\tau^\star, \zeta^{(t+1)} - \zeta_\tau^\star \right\rangle = 0. \tag{C.5}$$

By the definition of the KL divergence, one has

$$\begin{aligned}
&- \left\langle \log \zeta^{(t+1)} - (1-\eta\tau)\log \zeta^{(t)} - \eta\tau \log \zeta_\tau^\star, \zeta_\tau^\star \right\rangle \\
&= -(1-\eta\tau)\left\langle \log \zeta_\tau^\star - \log \zeta^{(t)}, \zeta_\tau^\star \right\rangle + \left\langle \log \zeta_\tau^\star - \log \zeta^{(t+1)}, \zeta_\tau^\star \right\rangle \\
&= -(1-\eta\tau)\mathsf{KL}\left(\zeta_\tau^\star \,\|\, \zeta^{(t)}\right) + \mathsf{KL}\left(\zeta_\tau^\star \,\|\, \zeta^{(t+1)}\right),
\end{aligned} \tag{C.6}$$

and similarly,

$$\begin{aligned}
&\left\langle \log \zeta^{(t+1)} - (1-\eta\tau)\log \zeta^{(t)} - \eta\tau \log \zeta_\tau^\star, \zeta^{(t+1)} \right\rangle \\
&= (1-\eta\tau)\left\langle \log \zeta^{(t+1)} - \log \zeta^{(t)}, \zeta^{(t+1)} \right\rangle + \eta\tau \left\langle \log \zeta^{(t+1)} - \log \zeta_\tau^\star, \zeta^{(t+1)} \right\rangle \\
&= (1-\eta\tau)\mathsf{KL}\left(\zeta^{(t+1)} \,\|\, \zeta^{(t)}\right) + \eta\tau\mathsf{KL}\left(\zeta^{(t+1)} \,\|\, \zeta_\tau^\star\right).
\end{aligned}$$

Combining the above two equalities with (C.5), we arrive at

$$\mathsf{KL}\left(\zeta_\tau^\star \,\|\, \zeta^{(t+1)}\right) + \eta\tau\mathsf{KL}\left(\zeta^{(t+1)} \,\|\, \zeta_\tau^\star\right) + (1-\eta\tau)\mathsf{KL}\left(\zeta^{(t+1)} \,\|\, \zeta^{(t)}\right) = (1-\eta\tau)\mathsf{KL}\left(\zeta_\tau^\star \,\|\, \zeta^{(t)}\right). \tag{C.7}$$

This immediately leads to $\mathsf{KL}\left(\zeta_\tau^\star \,\|\, \zeta^{(t+1)}\right) \le (1-\eta\tau)\mathsf{KL}\left(\zeta_\tau^\star \,\|\, \zeta^{(t)}\right)$ by the nonnegativity of the KL divergence, as long as $1-\eta\tau \ge 0$. Therefore

$$\mathsf{KL}\left(\zeta_\tau^\star \,\|\, \zeta^{(t)}\right) \le (1-\eta\tau)^t \mathsf{KL}\left(\zeta_\tau^\star \,\|\, \zeta^{(0)}\right) \qquad \text{for all } t \ge 0.$$

C.1.2 Proof of Theorem 5

Proof of policy convergence in KL divergence (4.9a)

First noticing that both PU and OMWU share the same update rule for $\mu^{(t+1)}$ and $\nu^{(t+1)}$, which takes the form

$$\begin{cases} \mu^{(t+1)}(a) \propto \mu^{(t)}(a)^{1-\eta\tau} \exp(\eta[A\bar{\nu}^{(t+1)}]_a), \\ \nu^{(t+1)}(b) \propto \nu^{(t)}(b)^{1-\eta\tau} \exp(-\eta[A^\top \bar{\mu}^{(t+1)}]_b). \end{cases}$$

Regarding this sequence, Lemma 13 (cf. (C.3)) gives

$$\left\langle \log \zeta^{(t+1)} - (1-\eta\tau)\log \zeta^{(t)} - \eta\tau \log \zeta_\tau^\star, \bar{\zeta}^{(t+1)} - \zeta_\tau^\star \right\rangle = 0. \tag{C.8}$$

In view of the similarity of (C.5) and (C.8), we can expect similar convergence guarantees to that of the implicit updates established in Proposition 1 with the optimism that $\bar{\zeta}^{(t+1)}$ approximates $\zeta^{(t+1)}$ well. Following the same argument as (C.6), we have

$$- \left\langle \log \zeta^{(t+1)} - (1-\eta\tau)\log \zeta^{(t)} - \eta\tau \log \zeta_\tau^\star, \zeta_\tau^\star \right\rangle = -(1-\eta\tau)\mathsf{KL}\left(\zeta_\tau^\star \,\|\, \zeta^{(t)}\right) + \mathsf{KL}\left(\zeta_\tau^\star \,\|\, \zeta^{(t+1)}\right). \tag{C.9}$$

On the other hand, it is easily seen that

$$\begin{aligned}
&\left\langle \log \zeta^{(t+1)} - (1-\eta\tau)\log \zeta^{(t)} - \eta\tau \log \zeta_\tau^\star, \bar{\zeta}^{(t+1)} \right\rangle \\
&= \left\langle \log \bar{\zeta}^{(t+1)} - (1-\eta\tau)\log \zeta^{(t)} - \eta\tau \log \zeta_\tau^\star, \bar{\zeta}^{(t+1)} \right\rangle - \left\langle \log \bar{\zeta}^{(t+1)} - \log \zeta^{(t+1)}, \zeta^{(t+1)} \right\rangle \\
&\qquad - \left\langle \log \bar{\zeta}^{(t+1)} - \log \zeta^{(t+1)}, \bar{\zeta}^{(t+1)} - \zeta^{(t+1)} \right\rangle \\
&= (1-\eta\tau)\mathsf{KL}\left(\bar{\zeta}^{(t+1)} \,\|\, \zeta^{(t)}\right) + \eta\tau \mathsf{KL}\left(\bar{\zeta}^{(t+1)} \,\|\, \zeta_\tau^\star\right) + \mathsf{KL}\left(\zeta^{(t+1)} \,\|\, \bar{\zeta}^{(t+1)}\right) \\
&\qquad - \left\langle \log \bar{\zeta}^{(t+1)} - \log \zeta^{(t+1)}, \bar{\zeta}^{(t+1)} - \zeta^{(t+1)} \right\rangle. \qquad \text{(C.10)}
\end{aligned}$$

Combining equalities (C.9), (C.10) with (C.8), we are left with the following relation pertaining to bounding $\mathsf{KL}\big(\zeta_\tau^\star \,\|\, \zeta^{(t)}\big)$:

$$\begin{aligned}
(1-\eta\tau)\mathsf{KL}\left(\zeta_\tau^\star \,\|\, \zeta^{(t)}\right) &= (1-\eta\tau)\mathsf{KL}\left(\bar{\zeta}^{(t+1)} \,\|\, \zeta^{(t)}\right) + \eta\tau \mathsf{KL}\left(\bar{\zeta}^{(t+1)} \,\|\, \zeta_\tau^\star\right) + \mathsf{KL}\left(\zeta^{(t+1)} \,\|\, \bar{\zeta}^{(t+1)}\right) \\
&\quad - \left\langle \log \bar{\zeta}^{(t+1)} - \log \zeta^{(t+1)}, \bar{\zeta}^{(t+1)} - \zeta^{(t+1)} \right\rangle + \mathsf{KL}\left(\zeta_\tau^\star \,\|\, \zeta^{(t+1)}\right). \qquad \text{(C.11)}
\end{aligned}$$

In addition, to bound $\mathsf{KL}\left(\zeta_\tau^\star \,\|\, \bar{\zeta}^{(t+1)}\right)$, we will resort to the following three-point equality, which reads

$$\begin{aligned}
\mathsf{KL}\left(\zeta_\tau^\star \,\|\, \bar{\zeta}^{(t+1)}\right) &= \mathsf{KL}\left(\zeta_\tau^\star \,\|\, \zeta^{(t+1)}\right) - \left\langle \zeta_\tau^\star, \log \bar{\zeta}^{(t+1)} - \log \zeta^{(t+1)} \right\rangle \\
&= \mathsf{KL}\left(\zeta_\tau^\star \,\|\, \zeta^{(t+1)}\right) - \mathsf{KL}\left(\bar{\zeta}^{(t+1)} \,\|\, \zeta^{(t+1)}\right) - \left\langle \zeta_\tau^\star - \bar{\zeta}^{(t+1)}, \log \bar{\zeta}^{(t+1)} - \log \zeta^{(t+1)} \right\rangle, \qquad \text{(C.12)}
\end{aligned}$$

which can be checked directly using the definition of the KL divergence.

To proceed, we need to control $\left\langle \log \bar{\zeta}^{(t+1)} - \log \zeta^{(t+1)}, \bar{\zeta}^{(t+1)} - \zeta^{(t+1)} \right\rangle$ on the right-hand side of inequality (C.11), and $\left\langle \zeta_\tau^\star - \bar{\zeta}^{(t+1)}, \log \bar{\zeta}^{(t+1)} - \log \zeta^{(t+1)} \right\rangle$ on the right-hand side of inequality (C.12), for which we continue the proofs for PU and OMWU separately as follows.

Bounding $\mathsf{KL}\big(\zeta_\tau^\star \,\|\, \zeta^{(t)}\big)$ for PU. Following the update rule of $\bar{\zeta}^{(t+1)} = (\bar{\mu}^{(t+1)}, \bar{\nu}^{(t+1)})$ in PU, we have

$$\log \bar{\mu}^{(t+1)} - \log \mu^{(t+1)} = \eta A(\nu^{(t)} - \bar{\nu}^{(t+1)}) + c \cdot \mathbf{1} \qquad \text{(C.13)}$$

for some normalization constant c. With this relation in place, one has

$$\begin{aligned}
\left\langle \log \bar{\mu}^{(t+1)} - \log \mu^{(t+1)}, \bar{\mu}^{(t+1)} - \mu^{(t+1)} \right\rangle &= \eta(\bar{\mu}^{(t+1)} - \mu^{(t+1)})^\top A(\nu^{(t)} - \bar{\nu}^{(t+1)}) \\
&\le \eta \|A\|_\infty \left\|\bar{\mu}^{(t+1)} - \mu^{(t+1)}\right\|_1 \left\|\bar{\nu}^{(t+1)} - \nu^{(t)}\right\|_1.
\end{aligned}$$

Combined with Pinsker's inequality, it is therefore clear that

$$\begin{aligned}
\left\langle \log \bar{\mu}^{(t+1)} - \log \mu^{(t+1)}, \bar{\mu}^{(t+1)} - \mu^{(t+1)} \right\rangle &\le \frac{1}{2}\eta \|A\|_\infty \left(\left\|\bar{\mu}^{(t+1)} - \mu^{(t+1)}\right\|_1^2 + \left\|\bar{\nu}^{(t+1)} - \nu^{(t)}\right\|_1^2\right) \\
&\le \eta \|A\|_\infty \left(\mathsf{KL}\left(\mu^{(t+1)} \,\|\, \bar{\mu}^{(t+1)}\right) + \mathsf{KL}\left(\bar{\nu}^{(t+1)} \,\|\, \nu^{(t)}\right)\right). \qquad \text{(C.14)}
\end{aligned}$$

Analogously, one can achieve the same bound regarding the quantity $\left\langle \log \bar{\nu}^{(t+1)} - \log \nu^{(t+1)}, \bar{\nu}^{(t+1)} - \nu^{(t+1)} \right\rangle$. Summing up these two inequalities, we end up with

$$\left\langle \log \bar{\zeta}^{(t+1)} - \log \zeta^{(t+1)}, \bar{\zeta}^{(t+1)} - \zeta^{(t+1)} \right\rangle \le \eta \left\| A \right\|_\infty \left(\mathsf{KL}\left(\zeta^{(t+1)} \,\|\, \bar{\zeta}^{(t+1)} \right) + \mathsf{KL}\left(\bar{\zeta}^{(t+1)} \,\|\, \zeta^{(t)} \right) \right).$$

Plugging the above inequality into inequality (C.11) leads to

$$\begin{aligned} \mathsf{KL}\left(\zeta_\tau^\star \,\|\, \zeta^{(t+1)} \right) \le (1-\eta\tau)\mathsf{KL}\left(\zeta_\tau^\star \,\|\, \zeta^{(t)} \right) - (1 - \eta\tau - \eta \left\| A \right\|_\infty)\mathsf{KL}\left(\bar{\zeta}^{(t+1)} \,\|\, \zeta^{(t)} \right) - \eta\tau \mathsf{KL}\left(\bar{\zeta}^{(t+1)} \,\|\, \zeta_\tau^\star \right) \\ - (1 - \eta \left\| A \right\|_\infty)\mathsf{KL}\left(\zeta^{(t+1)} \,\|\, \bar{\zeta}^{(t+1)} \right). \end{aligned} \tag{C.15}$$

Therefore, as long as the learning rate η satisfies $\eta \le \frac{1}{\tau + \|A\|_\infty}$, we are ensured that

$$\mathsf{KL}\left(\zeta_\tau^\star \,\|\, \zeta^{(t+1)} \right) \le (1-\eta\tau)\mathsf{KL}\left(\zeta_\tau^\star \,\|\, \zeta^{(t)} \right),$$

which further implies inequality (4.9a) when applied recursively.

Bounding $\mathsf{KL}\left(\zeta_\tau^\star \,\|\, \bar{\zeta}^{(t+1)} \right)$ for PU. By similar tricks of arriving at (C.14), we have

$$\begin{aligned} -\left\langle \mu_\tau^\star - \bar{\mu}^{(t+1)}, \log \bar{\mu}^{(t+1)} - \log \mu^{(t+1)} \right\rangle &= -\eta(\mu_\tau^\star - \bar{\mu}^{(t+1)})^\top A (\nu^{(t)} - \bar{\nu}^{(t+1)}) \\ &\le \frac{1}{2}\eta \left\| A \right\|_\infty \left(\left\| \mu_\tau^\star - \bar{\mu}^{(t+1)} \right\|_1^2 + \left\| \nu^{(t)} - \bar{\nu}^{(t+1)} \right\|_1^2 \right) \\ &\le \eta \left\| A \right\|_\infty \left(\mathsf{KL}\left(\mu_\tau^\star \,\|\, \bar{\mu}^{(t+1)} \right) + \mathsf{KL}\left(\bar{\nu}^{(t+1)} \,\|\, \nu^{(t)} \right) \right), \end{aligned}$$

following from (C.13) and Pinsker's inequality. A similar inequality for $-\left\langle \nu_\tau^\star - \bar{\nu}^{(t+1)}, \log \bar{\nu}^{(t+1)} - \log \nu^{(t+1)} \right\rangle$ can be obtained by symmetry, and summing together the two leads to

$$-\left\langle \zeta_\tau^\star - \bar{\zeta}^{(t+1)}, \log \bar{\zeta}^{(t+1)} - \log \zeta^{(t+1)} \right\rangle \le \eta \left\| A \right\|_\infty \left(\mathsf{KL}\left(\zeta_\tau^\star \,\|\, \bar{\zeta}^{(t+1)} \right) + \mathsf{KL}\left(\bar{\zeta}^{(t+1)} \,\|\, \zeta^{(t)} \right) \right).$$

Plugging the above inequality into (C.12) and rearranging terms, we reach at

$$(1 - \eta \left\| A \right\|_\infty)\mathsf{KL}\left(\zeta_\tau^\star \,\|\, \bar{\zeta}^{(t+1)} \right) \le \mathsf{KL}\left(\zeta_\tau^\star \,\|\, \zeta^{(t+1)} \right) + \eta \left\| A \right\|_\infty \mathsf{KL}\left(\bar{\zeta}^{(t+1)} \,\|\, \zeta^{(t)} \right).$$

Along with (C.15), we have

$$\begin{aligned} (1 - \eta \left\| A \right\|_\infty)\mathsf{KL}\left(\zeta_\tau^\star \,\|\, \bar{\zeta}^{(t+1)} \right) \le (1-\eta\tau)\mathsf{KL}\left(\zeta_\tau^\star \,\|\, \zeta^{(t)} \right) - (1 - \eta\tau - 2\eta \left\| A \right\|_\infty)\mathsf{KL}\left(\bar{\zeta}^{(t+1)} \,\|\, \zeta^{(t)} \right) \\ - \eta\tau \mathsf{KL}\left(\bar{\zeta}^{(t+1)} \,\|\, \zeta^\star \right) - (1 - \eta \left\| A \right\|_\infty)\mathsf{KL}\left(\zeta^{(t+1)} \,\|\, \bar{\zeta}^{(t+1)} \right). \end{aligned} \tag{C.16}$$

Therefore, with $\eta \le 1/(\tau + 2\left\| A \right\|_\infty)$ we have

$$\mathsf{KL}\left(\zeta_\tau^\star \,\|\, \bar{\zeta}^{(t+1)} \right) \le 2\mathsf{KL}\left(\zeta_\tau^\star \,\|\, \zeta^{(t)} \right) \le 2(1-\eta\tau)^t \mathsf{KL}\left(\zeta_\tau^\star \,\|\, \zeta^{(0)} \right).$$

Bounding $\mathsf{KL}\big(\zeta_\tau^\star \,\|\, \zeta^{(t)}\big)$ for OMWU. Following the update rule of $\bar{\zeta}^{(t+1)} = (\bar{\mu}^{(t+1)}, \bar{\nu}^{(t+1)})$ for OMWU, we have

$$\begin{aligned}\log \bar{\mu}^{(t+1)} - \log \mu^{(t+1)} &= \eta A(\bar{\nu}^{(t)} - \bar{\nu}^{(t+1)}) + c \cdot \mathbf{1} \\ &= \eta A(\bar{\nu}^{(t)} - \nu^{(t)}) + \eta A(\nu^{(t)} - \bar{\nu}^{(t+1)}) + c \cdot \mathbf{1}, \end{aligned} \tag{C.17}$$

where c is some normalization constant. Similar to the proof of relation (C.14), it can be easily demonstrated that

$$\begin{aligned}&\Big\langle \log \bar{\mu}^{(t+1)} - \log \mu^{(t+1)}, \bar{\mu}^{(t+1)} - \mu^{(t+1)} \Big\rangle \\ &= \eta(\bar{\mu}^{(t+1)} - \mu^{(t+1)})^\top A(\bar{\nu}^{(t)} - \nu^{(t)}) + \eta(\bar{\mu}^{(t+1)} - \mu^{(t+1)})^\top A(\nu^{(t)} - \bar{\nu}^{(t+1)}) \\ &\le \eta \left\|A\right\|_\infty \Big(\mathsf{KL}\left(\nu^{(t)} \,\|\, \bar{\nu}^{(t)}\right) + \mathsf{KL}\left(\bar{\nu}^{(t+1)} \,\|\, \nu^{(t)}\right) + 2\mathsf{KL}\left(\mu^{(t+1)} \,\|\, \bar{\mu}^{(t+1)}\right)\Big). \end{aligned} \tag{C.18}$$

By symmetry, we can also establish a similar inequality for $\big\langle \log \bar{\nu}^{(t+1)} - \log \nu^{(t+1)}, \bar{\nu}^{(t+1)} - \nu^{(t+1)} \big\rangle$, which in turns yields

$$\begin{aligned}&\Big\langle \log \bar{\zeta}^{(t+1)} - \log \zeta^{(t+1)}, \bar{\zeta}^{(t+1)} - \zeta^{(t+1)} \Big\rangle \\ &\le \eta \left\|A\right\|_\infty \Big(\mathsf{KL}\left(\zeta^{(t)} \,\|\, \bar{\zeta}^{(t)}\right) + \mathsf{KL}\left(\bar{\zeta}^{(t+1)} \,\|\, \zeta^{(t)}\right) + 2\mathsf{KL}\left(\zeta^{(t+1)} \,\|\, \bar{\zeta}^{(t+1)}\right)\Big). \end{aligned}$$

Plugging the above inequality into equation (C.11) and re-organizing terms, we arrive at

$$\begin{aligned}\mathsf{KL}\left(\zeta_\tau^\star \,\|\, \zeta^{(t+1)}\right) \le (1-\eta\tau)\mathsf{KL}\left(\zeta_\tau^\star \,\|\, \zeta^{(t)}\right) - (1-\eta\tau-\eta\left\|A\right\|_\infty)\mathsf{KL}\left(\bar{\zeta}^{(t+1)} \,\|\, \zeta^{(t)}\right) - \eta\tau\mathsf{KL}\left(\bar{\zeta}^{(t+1)} \,\|\, \zeta_\tau^\star\right) \\ - (1-2\eta\left\|A\right\|_\infty)\mathsf{KL}\left(\zeta^{(t+1)} \,\|\, \bar{\zeta}^{(t+1)}\right) + \eta\left\|A\right\|_\infty \mathsf{KL}\left(\zeta^{(t)} \,\|\, \bar{\zeta}^{(t)}\right). \end{aligned} \tag{C.19}$$

With the choice of the learning rate $\eta \le \min\{\frac{1}{2\|A\|_\infty + 2\tau}, \frac{1}{4\|A\|_\infty}\}$, it obeys

$$(1-\eta\tau)(1-2\eta\left\|A\right\|_\infty) \ge \eta\left\|A\right\|_\infty.$$

Combining the above inequality with (C.19) gives

$$\begin{aligned}&\mathsf{KL}\left(\zeta_\tau^\star \,\|\, \zeta^{(t+1)}\right) + (1-2\eta\left\|A\right\|_\infty)\mathsf{KL}\left(\zeta^{(t+1)} \,\|\, \bar{\zeta}^{(t+1)}\right) \\ &\le (1-\eta\tau)\mathsf{KL}\left(\zeta_\tau^\star \,\|\, \zeta^{(t)}\right) + \eta\left\|A\right\|_\infty \mathsf{KL}\left(\zeta^{(t)} \,\|\, \bar{\zeta}^{(t)}\right) - \eta\tau\mathsf{KL}\left(\bar{\zeta}^{(t+1)} \,\|\, \zeta_\tau^\star\right). \\ &\le (1-\eta\tau)\left[\mathsf{KL}\left(\zeta_\tau^\star \,\|\, \zeta^{(t)}\right) + (1-2\eta\left\|A\right\|_\infty)\mathsf{KL}\left(\zeta^{(t)} \,\|\, \bar{\zeta}^{(t)}\right)\right] - \eta\tau\mathsf{KL}\left(\bar{\zeta}^{(t+1)} \,\|\, \zeta_\tau^\star\right). \end{aligned}$$

For conciseness, let us introduce the shorthand notation

$$L^{(t)} := \mathsf{KL}\left(\zeta_\tau^\star \,\|\, \zeta^{(t)}\right) + (1-2\eta\left\|A\right\|_\infty)\mathsf{KL}\left(\zeta^{(t)} \,\|\, \bar{\zeta}^{(t)}\right). \tag{C.20}$$

As a result, the above inequality can be restated as

$$L^{(t+1)} \le (1-\eta\tau)L^{(t)} - \eta\tau\mathsf{KL}\left(\bar{\zeta}^{(t+1)} \,\|\, \zeta_\tau^\star\right). \tag{C.21}$$

Since we initialize OMWU with $\bar{\zeta}^{(0)} = \zeta^{(0)}$, therefore $L^{(0)} = \mathsf{KL}\left(\zeta_\tau^\star \,\|\, \zeta^{(0)}\right)$, which in turn gives

$$\mathsf{KL}\left(\zeta_\tau^\star \,\|\, \zeta^{(t)}\right) \le L^{(t)} \le (1-\eta\tau)^t L^{(0)} = (1-\eta\tau)^t \mathsf{KL}\left(\zeta_\tau^\star \,\|\, \zeta^{(0)}\right).$$

We complete the proof of inequality (4.9a) for OMWU.

Bounding $\mathsf{KL}\left(\zeta_\tau^\star \,\|\, \bar{\zeta}^{(t+1)}\right)$ for OMWU. By similar tricks of arriving at (C.18), we have

$$\begin{aligned}
&-\left\langle \mu_\tau^\star - \bar{\mu}^{(t+1)}, \log \bar{\mu}^{(t+1)} - \log \mu^{(t+1)} \right\rangle \\
&= \eta(\bar{\mu}^{(t+1)} - \mu_\tau^\star)^\top A(\bar{\nu}^{(t)} - \nu^{(t)}) + \eta(\bar{\mu}^{(t+1)} - \mu_\tau^\star)^\top A(\nu^{(t)} - \bar{\nu}^{(t+1)}) \\
&\le \eta \|A\|_\infty \left(\mathsf{KL}\left(\nu^{(t)} \,\|\, \bar{\nu}^{(t)}\right) + \mathsf{KL}\left(\bar{\nu}^{(t+1)} \,\|\, \nu^{(t)}\right) + 2\mathsf{KL}\left(\mu_\tau^\star \,\|\, \bar{\mu}^{(t+1)}\right)\right),
\end{aligned}$$

where the first line follows from (C.17). A similar inequality also holds for $-\left\langle \nu_\tau^\star - \bar{\nu}^{(t+1)}, \log \bar{\nu}^{(t+1)} - \log \nu^{(t+1)} \right\rangle$. Summing the two inequalities leads to

$$-\left\langle \zeta_\tau^\star - \bar{\zeta}^{(t+1)}, \log \bar{\zeta}^{(t+1)} - \log \zeta^{(t+1)} \right\rangle \le \eta \|A\|_\infty \left(\mathsf{KL}\left(\zeta^{(t)} \,\|\, \bar{\zeta}^{(t)}\right) + \mathsf{KL}\left(\bar{\zeta}^{(t+1)} \,\|\, \zeta^{(t)}\right) + 2\mathsf{KL}\left(\zeta_\tau^\star \,\|\, \bar{\zeta}^{(t+1)}\right)\right).$$

Plugging the above inequality into (C.12) and rearranging terms, we reach at

$$(1 - 2\eta\|A\|_\infty)\mathsf{KL}\left(\zeta_\tau^\star \,\|\, \bar{\zeta}^{(t+1)}\right) \le \mathsf{KL}\left(\zeta_\tau^\star \,\|\, \zeta^{(t+1)}\right) + \eta\|A\|_\infty \left(\mathsf{KL}\left(\zeta^{(t)} \,\|\, \bar{\zeta}^{(t)}\right) + \mathsf{KL}\left(\bar{\zeta}^{(t+1)} \,\|\, \zeta^{(t)}\right)\right).$$

Along with (C.19), we have

$$\begin{aligned}
&(1 - 2\eta\|A\|_\infty)\mathsf{KL}\left(\zeta_\tau^\star \,\|\, \bar{\zeta}^{(t+1)}\right) \\
&\le (1-\eta\tau)\mathsf{KL}\left(\zeta_\tau^\star \,\|\, \zeta^{(t)}\right) - (1 - \eta\tau - 2\eta\|A\|_\infty)\mathsf{KL}\left(\bar{\zeta}^{(t+1)} \,\|\, \zeta^{(t)}\right) - \eta\tau\mathsf{KL}\left(\bar{\zeta}^{(t+1)} \,\|\, \zeta_\tau^\star\right) \\
&\quad - (1 - 2\eta\|A\|_\infty)\mathsf{KL}\left(\zeta^{(t+1)} \,\|\, \bar{\zeta}^{(t+1)}\right) + 2\eta\|A\|_\infty \mathsf{KL}\left(\zeta^{(t)} \,\|\, \bar{\zeta}^{(t)}\right) \\
&\le (1-\eta\tau)\mathsf{KL}\left(\zeta_\tau^\star \,\|\, \zeta^{(t)}\right) + 2\eta\|A\|_\infty \mathsf{KL}\left(\zeta^{(t)} \,\|\, \bar{\zeta}^{(t)}\right) \\
&\le \mathsf{KL}\left(\zeta_\tau^\star \,\|\, \zeta^{(t)}\right) + (1 - 2\eta\|A\|_\infty)\mathsf{KL}\left(\zeta^{(t)} \,\|\, \bar{\zeta}^{(t)}\right) =: L^{(t)},
\end{aligned}$$

where we recall the shorthand notation $L^{(t)}$ in (C.20). As the learning rate of OMWU satisfies $0 < \eta < \min\left\{\frac{1}{2\|A\|_\infty + 2\tau}, \frac{1}{4\|A\|_\infty}\right\}$, it is clear that

$$\mathsf{KL}\left(\zeta_\tau^\star \,\|\, \bar{\zeta}^{(t+1)}\right) \le 2L^{(t)} \overset{\text{(i)}}{\le} 2(1-\eta\tau)^t L^{(0)} \le 2(1-\eta\tau)^t \mathsf{KL}\left(\zeta_\tau^\star \,\|\, \zeta^{(0)}\right),$$

where (i) follows from the recursive relation $L^{(t+1)} \le (1-\eta\tau)L^{(t)}$ shown in inequality (C.21).

Proof of entrywise convergence of policy log-ratios (4.9b)

To facilitate the proof, we introduce an auxiliary sequence $\{\xi^{(t)} \in \mathbb{R}^{|\mathcal{A}|}\}$ constructed recursively by

$$\xi^{(0)}(a) = \|\exp(A\nu_\tau^\star/\tau)\|_1 \cdot \mu^{(0)}(a), \tag{C.22a}$$

$$\xi^{(t+1)}(a) = \xi^{(t)}(a)^{1-\eta\tau} \exp(\eta[A\bar{\nu}^{(t+1)}]_a), \qquad \forall a \in \mathcal{A}, t \ge 0. \tag{C.22b}$$

It is easily seen that $\mu^{(t)}(a) \propto \xi^{(t)}(a) = \exp(\log \xi^{(t)}(a))$ for $t \ge 0$. Noticing that $\mu_\tau^\star \propto \exp(A\nu_\tau^\star)$, one has

$$\left\|\log \mu^{(t+1)} - \log \mu_\tau^\star\right\|_\infty \le 2\left\|\log \xi^{(t+1)} - A\nu_\tau^\star/\tau\right\|_\infty, \tag{C.23}$$

where we make use of Lemma 15.

Therefore it suffices for us to control the term $\left\|\log \xi^{(t+1)} - A\nu_\tau^\star/\tau\right\|_\infty$ on the right-hand side of inequality (C.23). Taking logarithm on both sides of (C.22b) yields

$$\begin{aligned}\log \xi^{(t+1)} - A\nu_\tau^\star/\tau &= (1-\eta\tau)\log \xi^{(t)} + \eta A\bar{\nu}^{(t+1)} - A\nu_\tau^\star/\tau \\ &= (1-\eta\tau)\left(\log \xi^{(t)} - A\nu_\tau^\star/\tau\right) + \eta A(\bar{\nu}^{(t+1)} - \nu_\tau^\star),\end{aligned}$$

which, when combined with Pinsker's inequality, implies

$$\begin{aligned}\left\|\log \xi^{(t+1)} - A\nu_\tau^\star/\tau\right\|_\infty &\le (1-\eta\tau)\left\|\log \xi^{(t)} - A\nu_\tau^\star/\tau\right\|_\infty + \eta\|A\|_\infty \left\|\bar{\nu}^{(t+1)} - \nu_\tau^\star\right\|_1 \\ &\le (1-\eta\tau)\left\|\log \xi^{(t)} - A\nu_\tau^\star/\tau\right\|_\infty + \eta\|A\|_\infty \left[2\mathsf{KL}\left(\nu_\tau^\star \,\|\, \bar{\nu}^{(t+1)}\right)\right]^{1/2} \\ &\le (1-\eta\tau)\left\|\log \xi^{(t)} - A\nu_\tau^\star/\tau\right\|_\infty + \eta\|A\|_\infty \left[2\mathsf{KL}\left(\zeta_\tau^\star \,\|\, \bar{\zeta}^{(t+1)}\right)\right]^{1/2}. \quad \text{(C.24)}\end{aligned}$$

Plugging the bound of $\mathsf{KL}\left(\zeta_\tau^\star \,\|\, \bar{\zeta}^{(t+1)}\right)$ from relation (4.9a) into (C.24) and invoking the inequality recursively leads to

$$\begin{aligned}&\left\|\log \xi^{(t+1)} - A\nu_\tau^\star/\tau\right\|_\infty \\ &\le (1-\eta\tau)^{t+1}\left\|\log \xi^{(0)} - A\nu_\tau^\star/\tau\right\|_\infty + 2\eta\|A\|_\infty \sum_{s=1}^{t+1}(1-\eta\tau)^{t+1-s/2}\mathsf{KL}\left(\zeta_\tau^\star \,\|\, \zeta^{(0)}\right)^{1/2} \\ &\le (1-\eta\tau)^{t+1}\left\|\log \xi^{(0)} - A\nu_\tau^\star/\tau\right\|_\infty + 2\eta\|A\|_\infty (1-\eta\tau)^{(t+1)/2}\frac{1}{1-(1-\eta\tau)^{1/2}}\mathsf{KL}\left(\zeta_\tau^\star \,\|\, \zeta^{(0)}\right)^{1/2} \\ &\le (1-\eta\tau)^{t+1}\left\|\log \xi^{(0)} - A\nu_\tau^\star/\tau\right\|_\infty + 4\tau^{-1}\|A\|_\infty (1-\eta\tau)^{(t+1)/2}\mathsf{KL}\left(\zeta_\tau^\star \,\|\, \zeta^{(0)}\right)^{1/2},\end{aligned}$$

where the last line results from the fact that $(1-\eta\tau)^{1/2} \le 1 - \eta\tau/2$. Combining pieces together, we end up with

$$\begin{aligned}\left\|\log \mu^{(t+1)} - \log \mu_\tau^\star\right\|_\infty &\le 2\left\|\log \xi^{(t+1)} - A\nu_\tau^\star/\tau\right\|_\infty \\ &\le 2(1-\eta\tau)^{t+1}\left\|\log \xi^{(0)} - A\nu_\tau^\star/\tau\right\|_\infty + 8\tau^{-1}\|A\|_\infty (1-\eta\tau)^{(t+1)/2}\mathsf{KL}\left(\zeta_\tau^\star \,\|\, \zeta^{(0)}\right)^{1/2} \\ &\le 2(1-\eta\tau)^{t+1}\left\|\log \mu^{(0)} - \log \mu_\tau^\star\right\|_\infty + 8\tau^{-1}\|A\|_\infty (1-\eta\tau)^{(t+1)/2}\mathsf{KL}\left(\zeta_\tau^\star \,\|\, \zeta^{(0)}\right)^{1/2}.\end{aligned}$$

Similarly, one can establish the corresponding inequality for $\left\|\log \nu^{(t+1)} - \log \nu_\tau^\star\right\|_\infty$, therefore completing the proof of inequality (4.9b).

Proof of convergence of optimality gap (4.9c)

To streamline our discussions, we only provide the proof of inequality (4.9c) concerning upper bounding $f_\tau(\bar{\mu}^{(t)}, \bar{\nu}^{(t)}) - f_\tau(\mu_\tau^\star, \nu_\tau^\star)$ without taking the absolute value; the other direction of the inequality can be established in the similar manner and hence is omitted.

We first make note of an important relation that holds both for PU and OMWU. Consider the update rule of $(\mu^{(t+1)}, \nu^{(t+1)})$, which is the same in PU and OMWU. Lemma 13 inequality (C.2a) gives

$$\left\langle \log \mu^{(t+1)} - (1-\eta\tau)\log \mu^{(t)} - \eta\tau \log \mu_\tau^\star, \bar{\mu}^{(t+1)} - \mu_\tau^\star \right\rangle = \eta(\mu_\tau^\star - \bar{\mu}^{(t+1)})^\top A(\nu_\tau^\star - \bar{\nu}^{(t+1)}). \quad \text{(C.25)}$$

Similar to what we have done in the proof of (4.9a) (cf. (C.11)), based on the above relation, we can therefore rearrange terms and conclude that

$$\begin{aligned}
&\eta\left(\tau\mathsf{KL}\left(\bar{\mu}^{(t+1)}\,\|\,\mu_\tau^\star\right)-(\mu_\tau^\star-\bar{\mu}^{(t+1)})^\top A(\nu_\tau^\star-\bar{\nu}^{(t+1)})\right)\\
&=(1-\eta\tau)\mathsf{KL}\left(\mu_\tau^\star\,\|\,\mu^{(t)}\right)-(1-\eta\tau)\mathsf{KL}\left(\bar{\mu}^{(t+1)}\,\|\,\mu^{(t)}\right)-\mathsf{KL}\left(\mu^{(t+1)}\,\|\,\bar{\mu}^{(t+1)}\right)\\
&\quad+\left\langle\log\bar{\mu}^{(t+1)}-\log\mu^{(t+1)},\bar{\mu}^{(t+1)}-\mu^{(t+1)}\right\rangle-\mathsf{KL}\left(\mu_\tau^\star\,\|\,\mu^{(t+1)}\right).
\end{aligned}\tag{C.26}$$

In conjunction with Lemma 14 (cf. (C.4b)), we can further derive

$$\begin{aligned}
&\eta\big(f_\tau(\mu_\tau^\star,\nu_\tau^\star)-f_\tau(\bar{\mu}^{(t+1)},\bar{\nu}^{(t+1)})\big)\le\eta\left(\tau\mathsf{KL}\left(\bar{\mu}^{(t+1)}\,\|\,\mu_\tau^\star\right)-(\mu_\tau^\star-\bar{\mu}^{(t+1)})^\top A(\nu_\tau^\star-\bar{\nu}^{(t+1)})\right)\\
&=(1-\eta\tau)\mathsf{KL}\left(\mu_\tau^\star\,\|\,\mu^{(t)}\right)-(1-\eta\tau)\mathsf{KL}\left(\bar{\mu}^{(t+1)}\,\|\,\mu^{(t)}\right)-\mathsf{KL}\left(\mu_\tau^\star\,\|\,\mu^{(t+1)}\right)\\
&\quad-\mathsf{KL}\left(\mu^{(t+1)}\,\|\,\bar{\mu}^{(t+1)}\right)+\left\langle\log\bar{\mu}^{(t+1)}-\log\mu^{(t+1)},\bar{\mu}^{(t+1)}-\mu^{(t+1)}\right\rangle,
\end{aligned}\tag{C.27}$$

where the second line follows from (C.26). From this point, we shall continue the proofs for PU and OMWU separately but follow similar strategies.

Remaining steps for PU. Plugging relation (C.14) into (C.27), we arrive at

$$\begin{aligned}
&\eta\big(f_\tau(\mu_\tau^\star,\nu_\tau^\star)-f_\tau(\bar{\mu}^{(t+1)},\bar{\nu}^{(t+1)})\big)\\
&\le(1-\eta\tau)\mathsf{KL}\left(\mu_\tau^\star\,\|\,\mu^{(t)}\right)-(1-\eta\tau)\mathsf{KL}\left(\bar{\mu}^{(t+1)}\,\|\,\mu^{(t)}\right)-\mathsf{KL}\left(\mu_\tau^\star\,\|\,\mu^{(t+1)}\right)\\
&\quad-(1-\eta\left\|A\right\|_\infty)\mathsf{KL}\left(\mu^{(t+1)}\,\|\,\bar{\mu}^{(t+1)}\right)+\eta\left\|A\right\|_\infty\mathsf{KL}\left(\bar{\nu}^{(t+1)}\,\|\,\nu^{(t)}\right)\\
&\le(1-\eta\tau)\mathsf{KL}\left(\mu_\tau^\star\,\|\,\mu^{(t)}\right)-\mathsf{KL}\left(\mu_\tau^\star\,\|\,\mu^{(t+1)}\right)-(1-\eta\tau)\mathsf{KL}\left(\bar{\mu}^{(t+1)}\,\|\,\mu^{(t)}\right)+\eta\left\|A\right\|_\infty\mathsf{KL}\left(\bar{\nu}^{(t+1)}\,\|\,\nu^{(t)}\right),
\end{aligned}\tag{C.28}$$

where the last line holds since $\eta(\tau+\left\|A\right\|_\infty)\le 1$. Similarly, from Lemma 13 inequality (C.2b), one can establish the following inequality in parallel

$$\begin{aligned}
&\eta\big(f_\tau(\bar{\mu}^{(t+1)},\bar{\nu}^{(t+1)})-f_\tau(\mu_\tau^\star,\nu_\tau^\star)\big)\\
&\le(1-\eta\tau)\mathsf{KL}\left(\nu_\tau^\star\,\|\,\nu^{(t)}\right)-\mathsf{KL}\left(\nu_\tau^\star\,\|\,\nu^{(t+1)}\right)-(1-\eta\tau)\mathsf{KL}\left(\bar{\nu}^{(t+1)}\,\|\,\nu^{(t)}\right)+\eta\left\|A\right\|_\infty\mathsf{KL}\left(\bar{\mu}^{(t+1)}\,\|\,\mu^{(t)}\right).
\end{aligned}\tag{C.29}$$

We are ready to establish inequality (4.9c) for PU. Computing (C.28) $+\frac{2}{3}\cdot$ (C.29) gives

$$\begin{aligned}
&\frac{\eta}{3}\big(f_\tau(\mu_\tau^\star,\nu_\tau^\star)-f_\tau(\bar{\mu}^{(t+1)},\bar{\nu}^{(t+1)})\big)\\
&\le(1-\eta\tau)\left[\mathsf{KL}\left(\mu_\tau^\star\,\|\,\mu^{(t)}\right)+\frac{2}{3}\mathsf{KL}\left(\nu_\tau^\star\,\|\,\nu^{(t)}\right)\right]-\left[\mathsf{KL}\left(\mu_\tau^\star\,\|\,\mu^{(t+1)}\right)+\frac{2}{3}\mathsf{KL}\left(\nu_\tau^\star\,\|\,\nu^{(t+1)}\right)\right]\\
&\quad-\left[(1-\eta\tau)-\frac{2}{3}\eta\left\|A\right\|_\infty\right]\mathsf{KL}\left(\bar{\mu}^{(t+1)}\,\|\,\mu^{(t)}\right)+\left[\eta\left\|A\right\|_\infty-\frac{2}{3}(1-\eta\tau)\right]\mathsf{KL}\left(\bar{\nu}^{(t+1)}\,\|\,\nu^{(t)}\right)\\
&\le(1-\eta\tau)\left[\mathsf{KL}\left(\mu_\tau^\star\,\|\,\mu^{(t)}\right)+\frac{2}{3}\mathsf{KL}\left(\nu_\tau^\star\,\|\,\nu^{(t)}\right)\right]-\left[\mathsf{KL}\left(\mu_\tau^\star\,\|\,\mu^{(t+1)}\right)+\frac{2}{3}\mathsf{KL}\left(\nu_\tau^\star\,\|\,\nu^{(t+1)}\right)\right]
\end{aligned}\tag{C.30}$$

Here, the last step is due to the fact that $(1-\eta\tau)-\frac{2}{3}\eta\left\|A\right\|_\infty \geq 0$ and $\eta\left\|A\right\|_\infty - \frac{2}{3}(1-\eta\tau)\leq 0$ when $0<\eta\leq \frac{1}{\tau+2\|A\|_\infty}$. As a direct consequence, the difference $f_\tau(\mu_\tau^\star,\nu_\tau^\star)-f_\tau(\bar{\mu}^{(t)},\bar{\nu}^{(t)})$ satisfies

$$\begin{aligned}
&\frac{\eta}{3}\big(f_\tau(\mu_\tau^\star,\nu_\tau^\star)-f_\tau(\bar{\mu}^{(t)},\bar{\nu}^{(t)})\big)\\
&\leq (1-\eta\tau)\left[(1-\eta\tau)\mathsf{KL}\left(\mu_\tau^\star\,\|\,\mu^{(t-1)}\right)+\eta\left\|A\right\|_\infty\mathsf{KL}\left(\nu_\tau^\star\,\|\,\nu^{(t-1)}\right)\right]\\
&\leq (1-\eta\tau)\mathsf{KL}\left(\zeta_\tau^\star\,\|\,\zeta^{(t-1)}\right)\leq (1-\eta\tau)^t\mathsf{KL}\left(\zeta_\tau^\star\,\|\,\zeta^{(0)}\right).
\end{aligned}$$

We conclude by noting that the other side of (4.9c) can be shown by considering $\frac{2}{3}\cdot$ (C.28) + (C.29) combined with similar arguments, and are therefore omitted.

Remaining steps for OMWU. Similar to the case of PU, plugging (C.18) into (C.27) gives

$$\begin{aligned}
&\eta\big(f_\tau(\mu_\tau^\star,\nu_\tau^\star)-f_\tau(\bar{\mu}^{(t+1)},\bar{\nu}^{(t+1)})\big)\\
&\leq (1-\eta\tau)\mathsf{KL}\left(\mu_\tau^\star\,\|\,\mu^{(t)}\right)-(1-\eta\tau)\mathsf{KL}\left(\bar{\mu}^{(t+1)}\,\|\,\mu^{(t)}\right)-\mathsf{KL}\left(\mu_\tau^\star\,\|\,\mu^{(t+1)}\right)\\
&\quad-(1-2\eta\left\|A\right\|_\infty)\mathsf{KL}\left(\mu^{(t+1)}\,\|\,\bar{\mu}^{(t+1)}\right)+\eta\left\|A\right\|_\infty\left[\mathsf{KL}\left(\nu^{(t)}\,\|\,\bar{\nu}^{(t)}\right)+\mathsf{KL}\left(\bar{\nu}^{(t+1)}\,\|\,\nu^{(t)}\right)\right]. \quad \text{(C.31)}
\end{aligned}$$

Similarly, one can establish a symmetric inequality as follows

$$\begin{aligned}
&\eta\big(f_\tau(\bar{\mu}^{(t+1)},\bar{\nu}^{(t+1)})-f_\tau(\mu_\tau^\star,\nu_\tau^\star)\big)\\
&\leq (1-\eta\tau)\mathsf{KL}\left(\nu_\tau^\star\,\|\,\nu^{(t)}\right)-(1-\eta\tau)\mathsf{KL}\left(\bar{\nu}^{(t+1)}\,\|\,\nu^{(t)}\right)-\mathsf{KL}\left(\nu_\tau^\star\,\|\,\nu^{(t+1)}\right)\\
&\quad-(1-2\eta\left\|A\right\|_\infty)\mathsf{KL}\left(\nu^{(t+1)}\,\|\,\bar{\nu}^{(t+1)}\right)+\eta\left\|A\right\|_\infty\left[\mathsf{KL}\left(\mu^{(t)}\,\|\,\bar{\mu}^{(t)}\right)+\mathsf{KL}\left(\bar{\mu}^{(t+1)}\,\|\,\mu^{(t)}\right)\right]. \quad \text{(C.32)}
\end{aligned}$$

Directly computing (C.31) $+\frac{2}{3}\cdot$ (C.32) gives

$$\begin{aligned}
&\frac{\eta}{3}\cdot\big(f_\tau(\mu_\tau^\star,\nu_\tau^\star)-f_\tau(\bar{\mu}^{(t+1)},\bar{\nu}^{(t+1)})\big)\\
&\leq (1-\eta\tau)\left[\mathsf{KL}\left(\mu_\tau^\star\,\|\,\mu^{(t)}\right)+\frac{2}{3}\mathsf{KL}\left(\nu_\tau^\star\,\|\,\nu^{(t)}\right)\right]-\left[\mathsf{KL}\left(\mu_\tau^\star\,\|\,\mu^{(t+1)}\right)+\frac{2}{3}\mathsf{KL}\left(\nu_\tau^\star\,\|\,\nu^{(t+1)}\right)\right]\\
&\quad-\left[(1-\eta\tau)-\frac{2}{3}\eta\left\|A\right\|_\infty\right]\mathsf{KL}\left(\bar{\mu}^{(t+1)}\,\|\,\mu^{(t)}\right)+\left[\eta\left\|A\right\|_\infty-\frac{2}{3}(1-\eta\tau)\right]\mathsf{KL}\left(\bar{\nu}^{(t+1)}\,\|\,\nu^{(t)}\right)\\
&\quad+\eta\left\|A\right\|_\infty\left[\frac{2}{3}\mathsf{KL}\left(\mu^{(t)}\,\|\,\bar{\mu}^{(t)}\right)+\mathsf{KL}\left(\nu^{(t)}\,\|\,\bar{\nu}^{(t)}\right)\right]\\
&\quad-(1-2\eta\left\|A\right\|_\infty)\left[\mathsf{KL}\left(\mu^{(t+1)}\,\|\,\bar{\mu}^{(t+1)}\right)+\frac{2}{3}\mathsf{KL}\left(\nu^{(t+1)}\,\|\,\bar{\nu}^{(t+1)}\right)\right]. \quad \text{(C.33)}
\end{aligned}$$

With our choice of the learning rate $\eta\leq\min\{\frac{1}{2\|A\|_\infty+2\tau},\frac{1}{4\|A\|_\infty}\}$, it is guarantees that

$$\eta\left\|A\right\|_\infty-\frac{2}{3}(1-\eta\tau)\leq 0,\quad (1-\eta\tau)-\frac{2}{3}\eta\left\|A\right\|_\infty\geq 0\quad\text{and}\quad (1-\eta\tau)(1-2\eta\left\|A\right\|_\infty)\geq\frac{3}{2}\eta\left\|A\right\|_\infty.$$

To proceed, let us introduce the shorthand notation

$$\begin{aligned}
G^{(t)}:=&\,\mathsf{KL}\left(\mu_\tau^\star\,\|\,\mu^{(t)}\right)+\frac{2}{3}\mathsf{KL}\left(\nu_\tau^\star\,\|\,\nu^{(t)}\right)\\
&+\frac{2}{3}(1-2\eta\left\|A\right\|_\infty)\left[\mathsf{KL}\left(\mu^{(t)}\,\|\,\bar{\mu}^{(t)}\right)+\mathsf{KL}\left(\nu^{(t)}\,\|\,\bar{\nu}^{(t)}\right)\right].
\end{aligned}$$

With this piece of notation, we can write inequality (C.33) as

$$\frac{\eta}{3}(f_\tau(\mu_\tau^\star, \nu_\tau^\star) - f_\tau(\bar{\mu}^{(t+1)}, \bar{\nu}^{(t+1)})) \leq (1-\eta\tau)G^{(t)} - G^{(t+1)}, \tag{C.34}$$

which in turn implies

$$\begin{aligned}&\frac{\eta}{3}(f_\tau(\mu_\tau^\star, \nu_\tau^\star) - f_\tau(\bar{\mu}^{(t)}, \bar{\nu}^{(t)}))\\ &\leq (1-\eta\tau)G^{(t-1)} \leq (1-\eta\tau)L^{(t-1)} \leq (1-\eta\tau)^t L^{(0)} = (1-\eta\tau)^t \mathsf{KL}\left(\zeta_\tau^\star \,\|\, \zeta^{(0)}\right),\end{aligned}$$

with $L^{(t)}$ defined in (C.20). This finishes the proof of (4.9c) for OMWU.

Proof of convergence of duality gap (4.9d)

The proof of inequality (4.9d) is built upon the following lemma whose proof is deferred to Appendix C.2.3.

Lemma 16. *The duality gap at $\zeta = (\mu, \nu)$ can be bounded as*

$$\max_{\mu' \in \Delta(\mathcal{A})} f_\tau(\mu', \nu) - \min_{\nu' \in \Delta(\mathcal{B})} f_\tau(\mu, \nu') \leq \tau \mathsf{KL}\left(\zeta \,\|\, \zeta_\tau^\star\right) + \tau^{-1}\|A\|_\infty^2 \mathsf{KL}\left(\zeta_\tau^\star \,\|\, \zeta\right).$$

Applying Lemma 16 to $\bar{\zeta}^{(t)} = (\bar{\mu}^{(t)}, \bar{\nu}^{(t)})$ yields

$$\begin{aligned}\mathsf{DualGap}_\tau(\bar{\zeta}^{(t)}) &\leq \tau \mathsf{KL}\left(\bar{\zeta}^{(t)} \,\|\, \zeta_\tau^\star\right) + \tau^{-1}\|A\|_\infty^2 \mathsf{KL}\left(\zeta_\tau^\star \,\|\, \bar{\zeta}^{(t)}\right)\\ &\leq \tau \mathsf{KL}\left(\bar{\zeta}^{(t)} \,\|\, \zeta_\tau^\star\right) + 2\tau^{-1}\|A\|_\infty^2 (1-\eta\tau)^{t-1} \mathsf{KL}\left(\zeta_\tau^\star \,\|\, \zeta^{(0)}\right),\end{aligned} \tag{C.35}$$

where the second step results from (4.9a). It remains to bound $\tau \mathsf{KL}\left(\bar{\zeta}^{(t)} \,\|\, \zeta_\tau^\star\right)$, which we proceed separately for PU and OMWU.

Remaining steps for PU. From inequality (C.15), we are ensured that

$$\eta\tau \mathsf{KL}\left(\bar{\zeta}^{(t)} \,\|\, \zeta_\tau^\star\right) \leq (1-\eta\tau)\mathsf{KL}\left(\zeta_\tau^\star \,\|\, \zeta^{(t-1)}\right) - \mathsf{KL}\left(\zeta_\tau^\star \,\|\, \zeta^{(t)}\right).$$

It thus follows that

$$\tau \mathsf{KL}\left(\bar{\zeta}^{(t)} \,\|\, \zeta_\tau^\star\right) \leq \eta^{-1}(1-\eta\tau)\mathsf{KL}\left(\zeta_\tau^\star \,\|\, \zeta^{(t-1)}\right) \leq \eta^{-1}(1-\eta\tau)^{t-1}\mathsf{KL}\left(\zeta_\tau^\star \,\|\, \zeta^{(0)}\right),$$

where the last inequality is due to inequality (4.9a). Plugging the above inequality into (C.35) completes the proof of inequality (4.9d) for PU.

Remaining steps for OMWU. From inequality (C.21), we are ensured that

$$\tau \mathsf{KL}\left(\bar{\zeta}^{(t)} \,\|\, \zeta_\tau^\star\right) \leq \eta^{-1}(1-\eta\tau)L^{(t-1)} \leq \eta^{-1}(1-\eta\tau)^t L^{(0)} = \eta^{-1}(1-\eta\tau)^t \mathsf{KL}\left(\zeta_\tau^\star \,\|\, \zeta^{(0)}\right),$$

where the last equality follows from $L^{(0)} = \mathsf{KL}\left(\zeta_\tau^\star \,\|\, \zeta^{(0)}\right)$. Plugging the above inequality into (C.35) finishes the proof of inequality (4.9d) for OMWU.

C.2 Proof of auxiliary lemmas

C.2.1 Proof of Lemma 13

Lemma 13 follows directly from the update sequence (4.6) and the form of the optimal solution pair $(\mu_\tau^\star, \nu_\tau^\star)$, provided in (4.4). Given the update sequence (4.6), taking logarithm of both sides of the first equation gives

$$\log \mu^{(t+1)} = (1-\eta\tau)\log \mu^{(t)} + \eta A z_2 + c \cdot \mathbf{1},$$

where c is the corresponding normalization constant. By rearranging terms and taking the inner product with $z_1 - \mu_\tau^\star$, we have

$$\left\langle \log \mu^{(t+1)} - (1-\eta\tau)\log \mu^{(t)}, z_1 - \mu_\tau^\star \right\rangle = \eta z_1^\top A z_2 - \eta {\mu_\tau^\star}^\top A z_2, \tag{C.36}$$

Similarly, one can derive

$$\left\langle \log \nu^{(t+1)} - (1-\eta\tau)\log \nu^{(t)}, z_2 - \nu_\tau^\star \right\rangle = -\eta z_1^\top A z_2 + \eta z_1^\top A \nu_\tau^\star. \tag{C.37}$$

By summing up equations (C.36) and (C.37), it is guarantee that

$$\left\langle \log \zeta^{(t+1)} - (1-\eta\tau)\log \zeta^{(t)}, \zeta_z - \zeta_\tau^\star \right\rangle = -\eta {\mu_\tau^\star}^\top A z_2 + \eta z_1^\top A \nu_\tau^\star, \tag{C.38}$$

where $\zeta(z) = (z_1, z_2)$.

On the other hand, recall the optimal policy pair $(\mu_\tau^\star, \nu_\tau^\star)$ satisfies the following fixed point equation

$$\begin{cases} \mu_\tau^\star(a) \propto \exp([A\nu_\tau^\star]_a/\tau), & \forall a \in \mathcal{A}, \\ \nu_\tau^\star(b) \propto \exp(-[A^\top \mu_\tau^\star]_b/\tau), & \forall b \in \mathcal{B}. \end{cases}$$

Taking logarithm of both sides of the first relation gives

$$\eta\tau \log \mu_\tau^\star = \eta A \nu_\tau^\star + c \cdot \mathbf{1}, \tag{C.39}$$

for some normalization constant c. Again, by taking the inner product with $z_1 - \mu_\tau^\star$, we have

$$\langle \eta\tau \log \mu_\tau^\star, z_1 - \mu_\tau^\star \rangle = \eta (z_1 - \mu_\tau^\star)^\top A \nu_\tau^\star, \tag{C.40}$$

and similarly

$$\langle \eta\tau \log \nu_\tau^\star, z_2 - \nu_\tau^\star \rangle = \eta {\mu_\tau^\star}^\top A (z_2 - \nu_\tau^\star). \tag{C.41}$$

Combining inequalities (C.36) and (C.40), we arrive at inequality (C.2a); combining inequalities (C.37) and (C.41) gives inequality (C.2b). Moreover, putting together inequalities (C.38), (C.40) and (C.41) leads to

$$\left\langle \log \zeta^{(t+1)} - (1-\eta\tau)\log \zeta^{(t)} - \eta\tau \log \zeta_\tau^\star, \zeta(z) - \zeta_\tau^\star \right\rangle = 0.$$

C.2.2 Proof of Lemma 14

We begin with establishing (C.4a). By the definition of $f_\tau(\mu, \nu)$, direct calculations yield

$$\begin{aligned} f_\tau(\mu_\tau^\star, \nu_\tau^\star) - f_\tau(\mu, \nu_\tau^\star) &= (\mu_\tau^\star - \mu)^\top A \nu_\tau^\star + \tau \mu^\top \log \mu - \tau {\mu_\tau^\star}^\top \log \mu_\tau^\star \\ &= \tau \left(\langle \mu_\tau^\star - \mu, \log \mu_\tau^\star \rangle + \mu^\top \log \mu - {\mu_\tau^\star}^\top \log \mu_\tau^\star \right) = \tau \mathsf{KL}\left(\mu \,\|\, \mu_\tau^\star\right). \end{aligned} \tag{C.42}$$

Here, the second equality is obtained by plugging in (C.39). Similarly, we have

$$f_\tau(\mu_\tau^\star, \nu) - f_\tau(\mu_\tau^\star, \nu_\tau^\star) = \tau \mathsf{KL}\left(\nu \,\|\, \nu_\tau^\star\right). \tag{C.43}$$

Summing these two equalities completes the proof of (C.4a).

Turning to (C.4b), we first write

$$\begin{aligned} f_\tau(\mu, \nu) + f_\tau(\mu_\tau^\star, \nu_\tau^\star) &= \mu^\top A \nu + {\mu_\tau^\star}^\top A \nu_\tau^\star + \tau \mathcal{H}(\mu) - \tau \mathcal{H}(\nu) + \tau \mathcal{H}(\mu_\tau^\star) - \tau \mathcal{H}(\nu_\tau^\star), \\ f_\tau(\mu_\tau^\star, \nu) + f_\tau(\mu, \nu_\tau^\star) &= {\mu_\tau^\star}^\top A \nu + \mu^\top A \nu_\tau^\star + \tau \mathcal{H}(\mu_\tau^\star) - \tau \mathcal{H}(\nu) + \tau \mathcal{H}(\mu) - \tau \mathcal{H}(\nu_\tau^\star). \end{aligned}$$

As a consequence, taking the difference of the above two equations leads to

$$f_\tau(\mu, \nu) + f_\tau(\mu_\tau^\star, \nu_\tau^\star) - f_\tau(\mu_\tau^\star, \nu) - f_\tau(\mu, \nu_\tau^\star) = (\mu_\tau^\star - \mu)^\top A (\nu_\tau^\star - \nu).$$

This in turn allows us to write $f_\tau(\mu, \nu) - f_\tau(\mu_\tau^\star, \nu_\tau^\star)$ as follows

$$f_\tau(\mu, \nu) - f_\tau(\mu_\tau^\star, \nu_\tau^\star) = (\mu_\tau^\star - \mu)^\top A (\nu_\tau^\star - \nu) + f_\tau(\mu_\tau^\star, \nu) + f_\tau(\mu, \nu_\tau^\star) - 2 f_\tau(\mu_\tau^\star, \nu_\tau^\star). \tag{C.44}$$

Finally, plugging (C.42) and (C.43) into (C.44) reveals the desired relation (C.4b).

C.2.3 Proof of Lemma 16

Since

$$\max_{\mu' \in \Delta(\mathcal{A})} f_\tau(\mu', \nu) - \min_{\nu' \in \Delta(\mathcal{B})} f_\tau(\mu, \nu') = \max_{\mu' \in \Delta(\mathcal{A}), \nu' \in \Delta(\mathcal{B})} f_\tau(\mu', \nu) - f_\tau(\mu, \nu'),$$

it boils down to control $f_\tau(\mu', \nu) - f_\tau(\mu, \nu')$ for any $(\mu', \nu') \in \Delta(\mathcal{A}) \times \Delta(\mathcal{B})$. Towards this, we have

$$\begin{aligned} f_\tau(\mu', \nu) - f_\tau(\mu, \nu') &= \left(f_\tau(\mu', \nu) - f_\tau(\mu', \nu_\tau^\star) - f_\tau(\mu, \nu') + f_\tau(\mu_\tau^\star, \nu')\right) - \left(f_\tau(\mu_\tau^\star, \nu') - f_\tau(\mu', \nu_\tau^\star)\right) \\ &= \left(f_\tau(\mu', \nu) - f_\tau(\mu', \nu_\tau^\star) - f_\tau(\mu, \nu') + f_\tau(\mu_\tau^\star, \nu')\right) - \tau \mathsf{KL}\left(\zeta' \,\|\, \zeta_\tau^\star\right), \end{aligned} \tag{C.45}$$

where the last step is due to $f_\tau(\mu, \nu_\tau^\star) - f_\tau(\mu_\tau^\star, \nu) = \tau \mathsf{KL}\left(\zeta \,\|\, \zeta_\tau^\star\right)$, as revealed in Lemma 14 (cf. (C.4a)).

To continue, observe that

$$\begin{aligned} f_\tau(\mu', \nu) - f_\tau(\mu', \nu_\tau^\star) &= {\mu'}^\top A (\nu - \nu_\tau^\star) + \nu^\top \log \nu - {\nu_\tau^\star}^\top \log \nu_\tau^\star \\ &= (\mu' - \mu_\tau^\star)^\top A (\nu - \nu_\tau^\star) + f_\tau(\mu_\tau^\star, \nu) - f_\tau(\mu_\tau^\star, \nu_\tau^\star). \end{aligned}$$

Similarly, we have

$$-f_\tau(\mu, \nu') + f_\tau(\mu_\tau^\star, \nu') = -\left(\mu - \mu_\tau^\star\right)^\top A (\nu' - \nu_\tau^\star) + f_\tau(\mu_\tau^\star, \nu_\tau^\star) - f_\tau(\mu, \nu_\tau^\star).$$

Plugging the above two equalities into (C.45) gives

$$\begin{aligned} & f_\tau(\mu', \nu) - f_\tau(\mu, \nu') \\ &= \left(\mu' - \mu_\tau^\star\right)^\top A (\nu - \nu_\tau^\star) - (\mu - \mu_\tau^\star)^\top A (\nu' - \nu_\tau^\star) + f_\tau(\mu_\tau^\star, \nu) - f_\tau(\mu, \nu_\tau^\star) - \tau \mathsf{KL}\left(\zeta' \,\|\, \zeta_\tau^\star\right) \\ &= \left(\mu' - \mu_\tau^\star\right)^\top A (\nu - \nu_\tau^\star) - (\mu - \mu_\tau^\star)^\top A (\nu' - \nu_\tau^\star) + \tau \mathsf{KL}\left(\zeta \,\|\, \zeta_\tau^\star\right) - \tau \mathsf{KL}\left(\zeta' \,\|\, \zeta_\tau^\star\right) \\ &\le \|A\|_\infty \left(\left\|\mu' - \mu_\tau^\star\right\|_1 \|\nu - \nu_\tau^\star\|_1 + \left\|\nu' - \nu_\tau^\star\right\|_1 \|\mu - \mu_\tau^\star\|_1\right) + \tau \mathsf{KL}\left(\zeta \,\|\, \zeta_\tau^\star\right) - \tau \mathsf{KL}\left(\zeta' \,\|\, \zeta_\tau^\star\right) \\ &\overset{\text{(i)}}{\le} \frac{1}{2} \|A\|_\infty \left[\frac{\tau}{\|A\|_\infty} \left(\left\|\mu' - \mu_\tau^\star\right\|_1^2 + \left\|\nu' - \nu_\tau^\star\right\|_1^2\right) + \frac{\|A\|_\infty}{\tau} \left(\|\mu - \mu_\tau^\star\|_1^2 + \|\nu - \nu_\tau^\star\|_1^2\right)\right] \\ &\qquad + \tau \mathsf{KL}\left(\zeta \,\|\, \zeta_\tau^\star\right) - \tau \mathsf{KL}\left(\zeta' \,\|\, \zeta_\tau^\star\right) \\ &\overset{\text{(ii)}}{\le} \tau \mathsf{KL}\left(\zeta' \,\|\, \zeta_\tau^\star\right) + \frac{\|A\|_\infty^2}{\tau} \mathsf{KL}\left(\zeta_\tau^\star \,\|\, \zeta\right) + \tau \mathsf{KL}\left(\zeta \,\|\, \zeta_\tau^\star\right) - \tau \mathsf{KL}\left(\zeta' \,\|\, \zeta_\tau^\star\right) \\ &= \frac{\|A\|_\infty^2}{\tau} \mathsf{KL}\left(\zeta_\tau^\star \,\|\, \zeta\right) + \tau \mathsf{KL}\left(\zeta \,\|\, \zeta_\tau^\star\right), \end{aligned}$$

where the second step invokes Lemma 14 (cf. (C.4a)), (i) follows from Young's inequality, namely $ab \le \frac{a^2}{2\varepsilon} + \frac{\varepsilon b^2}{2}$ with $\varepsilon = \frac{\|A\|_\infty}{\tau}$, and (ii) results from Pinsker's inequality. Taking maximum over μ', ν' finishes the proof.

Appendix D

Proofs for Chapter 5

D.1 Analysis for the infinite-horizon setting

We begin with the definitions of a certain concentrability coefficient, as well as the regularized minimax mismatch coefficient, which allow us to present general theorems that take into account the problem structure in a more refined manner, from which Theorem 6 follow directly.

Definition 4. *Given $\rho \in \Delta(\mathcal{S})$ with $\rho(s) > 0, \forall s \in S$, the concentrability coefficient $c_\rho(t)$ is defined as*

$$c_\rho(t) = \sup_{\substack{x^{(l)} \in \mathcal{A}^{\mathcal{S}}, 1 \le l \le t, \\ y^{(l)} \in \mathcal{B}^{\mathcal{S}}, 1 \le l \le t}} \left\| \frac{\rho P_{x^{(1)},y^{(1)}} \cdots P_{x^{(t)},y^{(t)}}}{\rho} \right\|_\infty,$$

where $P_{x^{(l)},y^{(l)}} \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{S}|}$ is the state transition matrix induced by a pair of deterministic policy $x^{(l)}, y^{(l)}$:

$$[P_{x^{(l)},y^{(l)}}]_{s,s'} = P(s'|s, x^{(l)}(s), y^{(l)}(s)).$$

Let $\mathcal{C}_\rho$ be the maximum value of $c_\rho(t)$ over $t \ge 0$:

$$\mathcal{C}_\rho = \sup_{t \ge 0} c_\rho(t).$$

In addition, let $\Gamma(\rho)$ be the set of all possible distribution over $\mathcal{S}$ induced by an initial state distribution ρ and deterministic policy sequences, i.e.,

$$\Gamma(\rho) = \bigcup_{t=0}^{\infty} \left\{ \rho P_{x^{(1)},y^{(1)}} \cdots P_{x^{(t)},y^{(t)}} : x^{(l)} \in \mathcal{A}^{\mathcal{S}}, y^{(l)} \in \mathcal{B}^{\mathcal{S}}, \forall l \in [t] \right\}.$$

The following definition of the the regularized minimax mismatch coefficient parallels that of the unregularized one in [Daskalakis et al., 2020].

Definition 5. *We define the regularized minimax mismatch coefficient by*

$$\mathcal{C}^\dagger_{\rho,\tau} = \max \left\{ \max_\mu \left\| \frac{d_\rho^{\mu,\nu^\dagger_\tau(\mu)}}{\rho} \right\|_\infty, \max_\nu \left\| \frac{d_\rho^{\mu^\dagger_\tau(\nu),\nu}}{\rho} \right\|_\infty \right\}.$$

Here, $\nu^\dagger_\tau(\mu)$ denotes the optimal policy of the min player when the max player adopts policy μ:

$$\nu^\dagger_\tau(\mu) = \arg\min_\nu V_\tau^{\mu,\nu}(\rho),$$

and $\mu_\tau^\dagger(\nu)$ is defined in a symmetric way. The discounted state visitation distribution $d_\rho^{\mu,\nu}$ is defined as

$$d_\rho^{\mu,\nu}(s) = (1-\gamma)\mathop{\mathbb{E}}_{s_0\sim\rho}\left[\sum_{t=0}^{\infty}\gamma^t P(s_t = s|s_0)\right].$$

We make note that Theorem 6 is the direct corollary of the following theorems, by setting ρ to the uniform distribution over $\mathcal{S}$, where $\mathcal{C}_\rho$ and $\|1/\rho\|_\infty$ admit a trivial upper bound $|\mathcal{S}|$. By a slight abuse of notation, let $\mathsf{KL}_\rho\big(\zeta\,\|\,\zeta'\big)$ denote $\mathbb{E}_{s\sim\rho}\big[\mathsf{KL}_s\big(\zeta\,\|\,\zeta'\big)\big]$ for $\rho\in\Delta(\mathcal{S})$.

Theorem 15. *With $0<\eta\le\frac{(1-\gamma)^3}{32000\mathcal{C}_\rho}$, and $\alpha_i=\eta\tau$, we have*

$$\max\left\{\mathsf{KL}_\rho\big(\zeta_\tau^\star\,\|\,\zeta^{(t)}\big),\frac{1}{2}\mathsf{KL}_\rho\big(\zeta_\tau^\star\,\|\,\bar\zeta^{(t)}\big),3\eta\mathop{\mathbb{E}}_{s\sim\rho}\left[\big\|Q^{(t)}(s)-Q_\tau^\star(s)\big\|_\infty\right]\right\}\le\frac{3000}{(1-\gamma)^2\tau}\Big(1-\frac{(1-\gamma)\eta\tau}{4}\Big)^t.$$

Theorem 16. *With $0<\eta\le\frac{(1-\gamma)^3}{32000\mathcal{C}_\rho}$, and $\alpha_i=\eta\tau$, we have*

$$\max_{s\in\mathcal{S},\mu,\nu}\Big(V_\tau^{\mu,\bar\nu^{(t)}}(s)-V_\tau^{\bar\mu^{(t)},\nu}(s)\Big)\le\frac{6000\|1/\rho\|_\infty}{(1-\gamma)^3\tau}\max\Big\{\frac{8}{(1-\gamma)^2\tau},\frac{1}{\eta}\Big\}\Big(1-\frac{(1-\gamma)\eta\tau}{4}\Big)^t,$$

and

$$\max_{\mu,\nu}\big(V_\tau^{\mu,\bar\nu^{(t)}}(\rho)-V_\tau^{\bar\mu^{(t)},\nu}(\rho)\big)\le\frac{6000\mathcal{C}_{\rho,\tau}^\dagger}{(1-\gamma)^3\tau}\max\Big\{\frac{8}{(1-\gamma)^2\tau},\frac{1}{\eta}\Big\}\Big(1-\frac{(1-\gamma)\eta\tau}{4}\Big)^t.$$

Key lemmas. While Theorem 15 and 16 focus on the case where $\alpha_i=\eta\tau,\forall i\ge1$, we assume in the following lemmas that the sequence $\{\alpha_i\}$ is non-increasing and bounded by $\eta\tau$ for generality. For notational simplicity, we set $Q^{(-1)}=0$, $\bar\zeta^{(-1)}=\bar\zeta^{(0)}$ and $\alpha_0=1$. It follows from the update rule (5.11a) that $\bar\zeta^{(1)}=\zeta^{(0)}=\bar\zeta^{(0)}$. Let us introduce

$$\alpha_{l,t}=\alpha_l\prod_{i=l+1}^{t}(1-\alpha_i),\tag{D.1}$$

and

$$\lambda_{l,t}=\alpha_l\prod_{i=l+1}^{t}\Big(1-\frac{1-\gamma}{4}\cdot\alpha_i\Big).\tag{D.2}$$

It follows straightforwardly that

$$\sum_{l=0}^{t}\alpha_{l,t}=\alpha_0=1.$$

We start with the following lemma.

Lemma 17. *Suppose $0<\eta\le1/\tau$. It holds for all $t\ge0$ that*

$$\begin{aligned}&\mathsf{KL}_\rho\big(\zeta_\tau^\star\,\|\,\zeta^{(t+1)}\big)-(1-\eta\tau)\mathsf{KL}_\rho\big(\zeta_\tau^\star\,\|\,\zeta^{(t)}\big)+\Big(1-\eta\tau-\frac{4\eta}{1-\gamma}\Big)\mathsf{KL}_\rho\big(\bar\zeta^{(t+1)}\,\|\,\bar\zeta^{(t)}\big)+\eta\tau\mathsf{KL}_\rho\big(\bar\zeta^{(t+1)}\,\|\,\zeta_\tau^\star\big)\\&\quad+\Big(1-\frac{2\eta}{1-\gamma}\Big)\mathsf{KL}_\rho\big(\zeta^{(t+1)}\,\|\,\bar\zeta^{(t+1)}\big)+(1-\eta\tau)\mathsf{KL}_\rho\big(\bar\zeta^{(t)}\,\|\,\zeta^{(t)}\big)-\frac{2\eta}{1-\gamma}\mathsf{KL}_\rho\big(\bar\zeta^{(t)}\,\|\,\bar\zeta^{(t-1)}\big)\\&\le\mathop{\mathbb{E}}_{s\sim\rho}\left[2\eta\big\|Q^{(t+1)}(s)-Q_\tau^\star(s)\big\|_\infty+\frac{4\eta^2}{1-\gamma}\big\|Q^{(t)}(s)-Q^{(t+1)}(s)\big\|_\infty+\frac{12\eta^2}{1-\gamma}\big\|Q^{(t-1)}(s)-Q^{(t)}(s)\big\|_\infty\right].\end{aligned}\tag{D.3}$$

Proof. See Appendix D.3.1. □

We continue to bound the terms on the right hand side of (D.3). By a slight abuse of notation, we denote

$$\big\|Q^{(t+1)}-Q_\tau^\star\big\|_{\Gamma(\rho)}=\sup_{\chi\in\Gamma(\rho)}\mathop{\mathbb{E}}_{s\sim\chi}\Big[\big\|Q^{(t+1)}(s)-Q_\tau^\star(s)\big\|_\infty\Big],$$

and

$$\big\|Q^{(t+1)}-Q^{(t)}\big\|_{\Gamma(\rho)}=\sup_{\chi\in\Gamma(\rho)}\mathop{\mathbb{E}}_{s\sim\chi}\Big[\big\|Q^{(t+1)}(s)-Q^{(t)}(s)\big\|_\infty\Big].$$

The following two lemmas establish a set of recursive bounds that relate $\big\{\big\|Q^{(l+1)}-Q_\tau^\star\big\|_{\Gamma(\rho)}\big\}_{0\le l\le t}$ and $\big\{\big\|Q^{(l+1)}-Q^{(l)}\big\|_{\Gamma(\rho)}\big\}_{0\le l\le t}$ with $\big\{\mathsf{KL}_\rho\big(\bar\zeta^{(l+1)}\,\|\,\bar\zeta^{(l)}\big)\big\}_{0\le l\le t-1}$.

Lemma 18. *Suppose that* $0<\eta\le\min\{(1-\gamma)/180,(1-\gamma)^2/48\}$. *It holds for all* $t\ge1$ *that*

$$\big\|Q^{(t+1)}-Q^{(t)}\big\|_{\Gamma(\rho)}\le\frac{1+\gamma}{2}\sum_{l=1}^{t}\alpha_{l,t}\big\|Q^{(l)}-Q^{(l-1)}\big\|_{\Gamma(\rho)}+\frac{4\mathcal{C}_\rho}{\eta}\cdot\sum_{l=1}^{t}\alpha_{l,t}\mathsf{KL}_\rho\big(\bar\zeta^{(l)}\,\|\,\bar\zeta^{(l-1)}\big).\tag{D.4}$$

When $t=0$, *we have* $\big\|Q^{(1)}-Q^{(0)}\big\|_{\Gamma(\rho)}\le2$.

Proof. See Appendix D.3.2. □

Lemma 19. *Suppose that* $0<\eta\le(1-\gamma)^2/16$. *It holds for all* $t\ge1$ *that*

$$\big\|Q^{(t+1)}-Q_\tau^\star\big\|_{\Gamma(\rho)}\le\frac{1+\gamma}{2}\cdot\sum_{l=0}^{t}\alpha_{l,t}\Big(\big\|Q^{(l)}-Q_\tau^\star\big\|_{\Gamma(\rho)}+\frac{2\eta}{1-\gamma}\big\|Q^{(l)}-Q^{(l-1)}\big\|_{\Gamma(\rho)}\Big)+2\alpha_{0,t}.\tag{D.5}$$

When $t=0$, *we have* $\big\|Q^{(1)}-Q_\tau^\star\big\|_{\Gamma(\rho)}\le\frac{2\gamma}{1-\gamma}$.

Proof. See Appendix D.3.3. □

The following lemma further demystifies the complicated recursive bounds showed in Lemmas 18-19.

Lemma 20. *Under the assumption of Lemma 18 and 19, it holds for all* $t\ge0$ *that*

$$\begin{aligned}&\sum_{l=0}^{t}\lambda_{l+1,t+1}\Big[\eta\big\|Q_\tau^\star-Q^{(l+1)}\big\|_{\Gamma(\rho)}+\frac{12\eta^2}{(1-\gamma)^2}\big\|Q^{(l+1)}-Q^{(l)}\big\|_{\Gamma(\rho)}\Big]\\&\le\frac{6250\eta\mathcal{C}_\rho}{(1-\gamma)^3}\sum_{l=0}^{t-1}\lambda_{l+1,t+1}\mathsf{KL}_\rho\big(\bar\zeta^{(l+1)}\,\|\,\bar\zeta^{(l)}\big)+\frac{550\eta}{(1-\gamma)^2}\lambda_{0,t+1}\end{aligned}$$

Proof. See Appendix D.3.4. □

Proof of Theorem 15. We are now ready to prove our main results. Starting with Lemma 17, averaging (D.3) with the weights $\lambda_{l,t}$ gives

$$\begin{aligned}
&\sum_{l=0}^{t}\lambda_{l+1,t+1}\Big[\mathsf{KL}_\rho\big(\zeta_\tau^\star\,\|\,\zeta^{(l+1)}\big)-(1-\eta\tau)\mathsf{KL}_\rho\big(\zeta_\tau^\star\,\|\,\zeta^{(l)}\big)\\
&\qquad+\Big(1-\frac{2\eta}{1-\gamma}\Big)\mathsf{KL}_\rho\big(\zeta^{(l+1)}\,\|\,\bar\zeta^{(l+1)}\big)+3\eta\mathop{\mathbb{E}}_{s\sim\rho}\Big[\big\|Q^{(l+1)}(s)-Q_\tau^\star(s)\big\|_\infty\Big]\\
&\qquad+\Big(1-\eta\tau-\frac{4\eta}{1-\gamma}\Big)\mathsf{KL}_\rho\big(\bar\zeta^{(l+1)}\,\|\,\bar\zeta^{(l)}\big)-\frac{2\eta}{1-\gamma}\mathsf{KL}_\rho\big(\bar\zeta^{(l)}\,\|\,\bar\zeta^{(l-1)}\big)\Big]\\
&\le\sum_{l=0}^{t}\lambda_{l+1,t+1}\mathop{\mathbb{E}}_{s\sim\rho}\Big[5\eta\big\|Q^{(l+1)}(s)-Q_\tau^\star(s)\big\|_\infty+\frac{4\eta^2}{1-\gamma}\big\|Q^{(l+1)}(s)-Q^{(l)}(s)\big\|_\infty\\
&\qquad+\frac{13\eta^2}{1-\gamma}\big\|Q^{(l-1)}(s)-Q^{(l)}(s)\big\|_\infty\Big]\\
&\le 5\sum_{l=0}^{t}\lambda_{l+1,t+1}\Big[\eta\big\|Q_\tau^\star-Q^{(l+1)}\big\|_{\Gamma(\rho)}+\frac{12\eta^2}{(1-\gamma)^2}\big\|Q^{(l+1)}-Q^{(l)}\big\|_{\Gamma(\rho)}\Big]\\
&\le\frac{31250\eta\mathcal{C}_\rho}{(1-\gamma)^3}\sum_{l=0}^{t-1}\lambda_{l+1,t+1}\mathsf{KL}_\rho\big(\bar\zeta^{(l+1)}\,\|\,\bar\zeta^{(l)}\big)+\frac{2750\eta}{(1-\gamma)^2}\lambda_{0,t+1}
\end{aligned}$$

for all $t\ge 0$, where the last line follows from Lemma 20. Rearranging terms, we have

$$\begin{aligned}
&\alpha_{t+1}\Big[\mathsf{KL}_\rho\big(\zeta_\tau^\star\,\|\,\zeta^{(t+1)}\big)+\Big(1-\frac{2\eta}{1-\gamma}\Big)\mathsf{KL}_\rho\big(\zeta^{(t+1)}\,\|\,\bar\zeta^{(t+1)}\big)+3\eta\mathop{\mathbb{E}}_{s\sim\rho}\Big[\big\|Q^{(t+1)}(s)-Q_\tau^\star(s)\big\|_\infty\Big]\Big]\\
&\quad+\sum_{l=1}^{t}(\lambda_{l,t+1}-(1-\eta\tau)\lambda_{l+1,t+1})\mathsf{KL}_\rho\big(\zeta_\tau^\star\,\|\,\zeta^{(l)}\big)\\
&\quad+\sum_{l=0}^{t-1}\Big[\lambda_{l+1,t+1}\Big(1-\eta\tau-\frac{4\eta}{1-\gamma}-\frac{31250\eta\mathcal{C}_\rho}{(1-\gamma)^3}\Big)-\lambda_{l+2,t+1}\frac{2\eta}{1-\gamma}\Big]\mathsf{KL}\Big(\bar\zeta^{(l+1)}\,\|\,\bar\zeta^{(l)}\Big)\\
&\le\frac{2750\eta}{(1-\gamma)^2}\lambda_{0,t+1}+(1-\eta\tau)\lambda_{1,t+1}\mathsf{KL}_\rho\big(\zeta_\tau^\star\,\|\,\zeta^{(0)}\big)\le\Big(\frac{2750\eta}{(1-\gamma)^2}+\eta\Big)\lambda_{0,t+1}.
\end{aligned}$$

Here, the last step results from

$$\begin{aligned}
(1-\eta\tau)\lambda_{1,t+1}\mathsf{KL}_\rho\big(\zeta_\tau^\star\,\|\,\zeta^{(0)}\big)&=\alpha_1\cdot\frac{1-\eta\tau}{1-(1-\gamma)\alpha_1/4}\lambda_{0,t+1}\mathsf{KL}_\rho\big(\zeta_\tau^\star\,\|\,\zeta^{(0)}\big)\\
&\le\eta\tau\lambda_{0,t+1}\mathsf{KL}_\rho\big(\zeta_\tau^\star\,\|\,\zeta^{(0)}\big)\le\eta\tau(\log|\mathcal{A}|+\log|\mathcal{B}|)\lambda_{0,t+1}\le\eta\lambda_{0,t+1}.
\end{aligned}$$

where we use the fact that $\alpha_1=\eta\tau$ and the assumption on τ (5.7). With $0<\eta\le\frac{(1-\gamma)^3}{32000\mathcal{C}_\rho}$, and $\alpha_i=\eta\tau$, we have $\lambda_{l,t+1}-(1-\eta\tau)\lambda_{l+1,t+1}\ge 0$ (cf. (D.33)), and

$$\begin{aligned}
&\lambda_{l+1,t+1}\Big(1-\eta\tau-\frac{4\eta}{1-\gamma}-\frac{31250\eta\mathcal{C}_\rho}{(1-\gamma)^3}\Big)-\lambda_{l+2,t+1}\frac{2\eta}{1-\gamma}\\
&=\eta\tau\prod_{j=l+3}^{t+1}\Big(1-\frac{1-\gamma}{4}\alpha_j\Big)\Big[(1-\frac{1-\gamma}{4}\eta\tau)\Big(1-\eta\tau-\frac{4\eta}{1-\gamma}-\frac{31250\eta\mathcal{C}_\rho}{(1-\gamma)^3}\Big)-\frac{2\eta}{1-\gamma}\Big]\ge 0.
\end{aligned}$$

It follows that

$$\begin{aligned}
&\mathsf{KL}_\rho\big(\zeta_\tau^\star \,\|\, \zeta^{(t+1)}\big) + \Big(1 - \frac{2\eta}{1-\gamma}\Big)\mathsf{KL}_\rho\big(\zeta^{(t+1)} \,\|\, \bar{\zeta}^{(t+1)}\big) + 3\eta \mathop{\mathbb{E}}_{s\sim\rho}\Big[\big\|Q^{(t+1)}(s) - Q_\tau^\star(s)\big\|_\infty\Big] \\
&\le \Big(\frac{2750}{(1-\gamma)^2\tau} + \frac{1}{\tau}\Big)\Big(1 - \frac{(1-\gamma)\eta\tau}{4}\Big)^{t+1} < \frac{3000}{(1-\gamma)^2\tau}\Big(1 - \frac{(1-\gamma)\eta\tau}{4}\Big)^{t+1}.
\end{aligned} \tag{D.6}$$

This proves the bound of $\mathsf{KL}_\rho\big(\zeta_\tau^\star \,\|\, \zeta^{(t+1)}\big)$ and $3\eta \mathop{\mathbb{E}}_{s\sim\rho}\big[\big\|Q^{(t+1)}(s) - Q_\tau^\star(s)\big\|_\infty\big]$ in Theorem 15. Note that the bound holds trivially for $\mathsf{KL}_\rho\big(\zeta_\tau^\star \,\|\, \zeta^{(0)}\big)$ and $3\eta \mathop{\mathbb{E}}_{s\sim\rho}\big[\big\|Q^{(0)}(s) - Q_\tau^\star(s)\big\|_\infty\big]$. It remains to bound $\mathsf{KL}\left(\zeta_\tau^\star \,\|\, \bar{\zeta}^{(t+1)}\right)$, which we make use of the following lemma.

Lemma 21. *With $0 < \eta \le (1-\gamma)/8$, we have*

$$\begin{aligned}
&\frac{1}{2}\mathsf{KL}_s\big(\zeta_\tau^\star \,\|\, \bar{\zeta}^{(t+1)}\big) + \eta\tau\mathsf{KL}_s\big(\bar{\zeta}^{(t+1)} \,\|\, \zeta_\tau^\star\big) \\
&\le (1-\eta\tau)\mathsf{KL}_s\big(\zeta_\tau^\star \,\|\, \zeta^{(t)}\big) + \frac{2\eta}{1-\gamma}\mathsf{KL}_s\big(\zeta^{(t)} \,\|\, \bar{\zeta}^{(t)}\big) + 2\eta\big\|Q^{(t)}(s) - Q_\tau^\star(s)\big\|_\infty.
\end{aligned}$$

Proof. See Appendix D.3.5. □

Combining Lemma 21 with (D.6) gives

$$\begin{aligned}
&\frac{1}{2}\mathsf{KL}_\rho\big(\zeta_\tau^\star \,\|\, \bar{\zeta}^{(t+1)}\big) + \eta\tau\mathsf{KL}_\rho\big(\bar{\zeta}^{(t+1)} \,\|\, \zeta_\tau^\star\big) \\
&\le (1-\eta\tau)\Big(\mathsf{KL}_\rho\big(\zeta_\tau^\star \,\|\, \zeta^{(t)}\big) + \Big(1 - \frac{2\eta}{1-\gamma}\Big)\mathsf{KL}_\rho\big(\zeta^{(t)} \,\|\, \bar{\zeta}^{(t)}\big) + 3\eta \mathop{\mathbb{E}}_{s\sim\rho}\Big[\big\|Q^{(t)}(s) - Q_\tau^\star(s)\big\|_\infty\Big]\Big) \\
&\le \frac{3000}{(1-\gamma)^2\tau}\Big(1 - \frac{(1-\gamma)\eta\tau}{4}\Big)^{t+1},
\end{aligned} \tag{D.7}$$

which concludes the proof of Theorem 15.

Proof of Theorem 16. We are now ready to bound the duality gap in Theorem 16. Before proceeding, we introduce the following two lemmas.

Lemma 22. *It holds for any policy pair (μ, ν) that*

$$\max_{\mu',\nu'}\Big(V_\tau^{\mu',\nu}(\rho) - V_\tau^{\mu,\nu'}(\rho)\Big) \le \frac{2\mathcal{C}_{\rho,\tau}^\dagger}{1-\gamma}\mathop{\mathbb{E}}_{s\sim\rho}\Big[\max_{\mu',\nu'}\Big(f_s(Q_\tau^\star, \mu', \nu) - f_s(Q_\tau^\star, \mu, \nu')\Big)\Big] \tag{D.8}$$

and

$$\max_{s\in\mathcal{S},\mu',\nu'}\Big(V_\tau^{\mu',\nu}(s) - V_\tau^{\mu,\nu'}(s)\Big) \le \frac{2\|1/\rho\|_\infty}{1-\gamma}\mathop{\mathbb{E}}_{s\sim\rho}\Big[\max_{\mu',\nu'}\Big(f_s(Q_\tau^\star, \mu', \nu) - f_s(Q_\tau^\star, \mu, \nu')\Big)\Big]. \tag{D.9}$$

Here, $f_s(Q, \mu, \nu)$ is the one-step entropy-regularized game value at state s, i.e.,

$$f_s(Q, \mu, \nu) = \mu(s)^\top Q(s)\nu(s) + \tau\mathcal{H}(\mu(s)) - \tau\mathcal{H}(\nu(s)). \tag{D.10}$$

Proof. Note that (D.9) is a slight generalization of [Wei et al., 2021b, Lemma 32]. The proof can be found in Appendix D.3.6. □

Lemma 23 ([Cen et al., 2021, Lemma 4]). *It holds for all $s \in \mathcal{S}$ and policy pair μ, ν that*

$$\max_{\mu',\nu'} \big(f_s(Q_\tau^\star, \mu', \nu) - f_s(Q_\tau^\star, \mu, \nu')\big) \le \frac{4}{(1-\gamma)^2\tau}\mathsf{KL}_s\big(\zeta_\tau^\star \,\|\, \zeta\big) + \tau\mathsf{KL}_s\big(\zeta \,\|\, \zeta_\tau^\star\big).$$

Putting all pieces together, we arrive at

$$\begin{aligned}
\max_{\mu,\nu}\big(V_\tau^{\mu,\bar{\nu}^{(t)}}(\rho) - V_\tau^{\bar{\mu}^{(t)},\nu}(\rho)\big) &\le \frac{2\mathcal{C}_{\rho,\tau}^\dagger}{1-\gamma}\Big(\frac{4}{(1-\gamma)^2\tau}\mathsf{KL}_\rho\big(\zeta_\tau^\star \,\|\, \bar{\zeta}^{(t+1)}\big) + \tau\mathsf{KL}_\rho\big(\bar{\zeta}^{(t+1)} \,\|\, \zeta_\tau^\star\big)\Big) \\
&\le \frac{2\mathcal{C}_{\rho,\tau}^\dagger}{1-\gamma}\max\Big\{\frac{8}{(1-\gamma)^2\tau}, \frac{1}{\eta}\Big\}\Big(\frac{1}{2}\mathsf{KL}_\rho\big(\zeta_\tau^\star \,\|\, \bar{\zeta}^{(t+1)}\big) + \eta\tau\mathsf{KL}_\rho\big(\bar{\zeta}^{(t+1)} \,\|\, \zeta_\tau^\star\big)\Big) \\
&\le \frac{6000\mathcal{C}_{\rho,\tau}^\dagger}{(1-\gamma)^3\tau}\max\Big\{\frac{8}{(1-\gamma)^2\tau}, \frac{1}{\eta}\Big\}\Big(1 - \frac{(1-\gamma)\eta\tau}{4}\Big)^t,
\end{aligned}$$

where the last line follows from (D.7). We omit the proof for $\max_{s\in\mathcal{S},\mu,\nu}\Big(V_\tau^{\mu,\bar{\nu}^{(t)}}(s) - V_\tau^{\bar{\mu}^{(t)},\nu}(s)\Big)$ for brevity as it follows essentially from the same argument.

D.2 Analysis for the finite-horizon setting

Throughout the analysis, we restrict our choice of the step size for value update to $\alpha_t = \eta\tau$. We start with the following lemma which parallels Lemma 27 in the infinite-horizon Markov game setting; for brevity we omit the proof.

Lemma 24. *With $0 < \eta \le 1/\tau$, it holds for all $s \in \mathcal{S}$, $h \in [H]$ and $t \ge 0$ that*

$$\max\Big\{\big\|\bar{\mu}_h^{(t+1)}(s) - \mu_h^{(t+1)}(s)\big\|_1, \big\|\bar{\nu}_h^{(t+1)}(s) - \nu_h^{(t+1)}(s)\big\|_1\Big\} \le 2\eta H. \tag{D.11}$$

In addition, we have

$$\max\{\|\log\zeta_h^{(t)}(s)\|_\infty, \|\log\bar{\zeta}_h^{(t)}(s)\|_\infty, \|\log\zeta_{h,\tau}^\star(s)\|_\infty\} \le \frac{2H}{\tau}. \tag{D.12}$$

Lemma 25. *With $0 < \eta \le \frac{1}{8H}$, it holds for all $0 \le t_1 \le t_2$, $h \in [H]$ and $s \in \mathcal{S}$ that*

$$\begin{aligned}
&\mathsf{KL}_s\big(\zeta_{h,\tau}^\star \,\|\, \zeta_h^{(t_2)}\big) + (1-4\eta H)\mathsf{KL}_s\big(\zeta_h^{(t_2)} \,\|\, \bar{\zeta}_h^{(t_2)}\big) \\
&\le (1-\eta\tau)^{t_2-t_1}\Big(\mathsf{KL}_s\big(\zeta_{h,\tau}^\star \,\|\, \zeta_h^{(t_1)}\big) + (1-4\eta H)\mathsf{KL}_s\big(\zeta_h^{(t_1)} \,\|\, \bar{\zeta}_h^{(t_1)}\big)\Big) + 4\eta\sum_{l=t_1}^{t_2}(1-\eta\tau)^{t_2-l}\big\|Q_h^{(l)}(s) - Q_\tau^\star(s)\big\|_\infty.
\end{aligned}$$

Proof. See Appendix D.4.1. □

Lemma 26. *With $0 < \eta \le \frac{1}{8H}$, it holds for all $0 < t_1 \le t_2$, $2 \le h \le H$ and $s \in \mathcal{S}$ that*

$$\begin{aligned}
&\big|Q_{h-1}^{(t_2)}(s,a,b) - Q_{h-1,\tau}^\star(s,a,b)\big| \\
&\le 2(1-\eta\tau)^{t_2-t_1}H + 10\eta\tau\mathop{\mathbb{E}}_{s'\sim P_{h-1}(\cdot|s,a,b)}\left[\sum_{l=t_1-1}^{t_2-1}(1-\eta\tau)^{t_2-1-l}\big\|Q_h^{(l)}(s) - Q_{h,\tau}^\star(s)\big\|_\infty\right] \\
&\quad + \tau(1-\eta\tau)^{t_2-t_1}\mathop{\mathbb{E}}_{s'\sim P_{h-1}(\cdot|s,a,b)}\Big[\mathsf{KL}_s\big(\zeta_{h,\tau}^\star \,\|\, \zeta_h^{(t_1-1)}\big) + (1-4\eta H)\mathsf{KL}_s\big(\zeta_h^{(t_1-1)} \,\|\, \bar{\zeta}_h^{(t_1-1)}\big)\Big].
\end{aligned}$$

Proof. See Appendix D.4.2. □

Proof of Theorem 7. We prove Theorem 7 by induction. By definition, we have

$$\big\|Q^\star_{H,\tau} - Q^{(0)}_H\big\|_\infty = \big\|Q^\star_{H,\tau}\big\|_\infty \le 1,$$

and $\big\|Q^\star_{H,\tau} - Q^{(t)}_H\big\|_\infty = \big\|r_H - r_H\big\|_\infty = 0$ for $t > 0$. So (5.16a) holds trivially for $h = H$. When the statement holds for some h, we can invoke Lemma 26 with $t_1 = T_h + 1$ and $t_2 = t \ge T_{h-1}$, which yields

$$\begin{aligned}
\big\|Q^{(t)}_{h-1} - Q^\star_{h-1,\tau}\big\| &\le 2(1-\eta\tau)^{t-T_h-1}H + 10\eta\tau \mathop{\mathbb{E}}_{s'\sim P(\cdot|s,a,b)}\left[\sum_{l=T_h}^{t-1}(1-\eta\tau)^{t-1-l}\big\|Q^{(l)}_h(s) - Q^\star_{h,\tau}(s)\big\|_\infty\right] \\
&\quad + \tau(1-\eta\tau)^{t-T_h-1}\mathop{\mathbb{E}}_{s'\sim P(\cdot|s,a,b)}\Big[\mathsf{KL}_s\big(\zeta^\star_{h,\tau}\,\|\,\zeta^{(T_h)}_h\big) + (1-4\eta H)\mathsf{KL}_s\big(\zeta^{(T_h)}_h\,\|\,\bar\zeta^{(T_h)}_h\big)\Big] \\
&\le 2(1-\eta\tau)^{t-T_h-1}H + 10\eta\tau \mathop{\mathbb{E}}_{s'\sim P(\cdot|s,a,b)}\left[\sum_{l=T_h}^{t-1}(1-\eta\tau)^{t-T_h-1}l^{H-h}\right] \\
&\quad + \tau(1-\eta\tau)^{t-T_h-1}\mathop{\mathbb{E}}_{s'\sim P(\cdot|s,a,b)}\Big[\mathsf{KL}_s\big(\zeta^\star_{h,\tau}\,\|\,\zeta^{(T_h)}_h\big) + (1-4\eta H)\mathsf{KL}_s\big(\zeta^{(T_h)}_h\,\|\,\bar\zeta^{(T_h)}_h\big)\Big] \\
&\le (1-\eta\tau)^{t-T_{h-1}}(1-\eta\tau)^{T_{\mathsf{start}}-1}\Big[10H + 10\eta\tau t^{H-h+1}\Big],
\end{aligned}$$

where the last step results from

$$\begin{aligned}
&\tau\Big(\mathsf{KL}_s\big(\zeta^\star_{h,\tau}\,\|\,\zeta^{(T_h)}_h\big) + (1-4\eta H)\mathsf{KL}_s\big(\zeta^{(T_h)}_h\,\|\,\bar\zeta^{(T_h)}_h\big)\Big) \\
&\le \tau\Big(\big\|\log\mu^\star_{h,\tau}(s) - \log\mu^{(T_h)}_h(s)\big\|_\infty + \big\|\log\nu^\star_{h,\tau}(s) - \log\nu^{(T_h)}_h(s)\big\|_\infty \\
&\qquad + \big\|\log\mu^{(T_h)}_h(s) - \log\bar\mu^{(T_h)}_h(s)\big\|_\infty + \big\|\log\nu^{(T_h)}_h(s) - \log\bar\nu^{(T_h)}_h(s)\big\|_\infty\Big) \\
&\le \tau\Big(\max\big\{\big\|\log\mu^\star_{h,\tau}(s)\big\|_\infty\big\|\log\mu^{(T_h)}_h(s)\big\|_\infty\big\} + \max\big\{\big\|\log\nu^\star_{h,\tau}(s)\big\|_\infty, \big\|\log\nu^{(T_h)}_h(s)\big\|_\infty\big\} \\
&\qquad + \max\big\{\big\|\log\mu^{(T_h)}_h(s)\big\|_\infty, \big\|\log\bar\mu^{(T_h)}_h(s)\big\|_\infty\big\} + \max\big\{\big\|\log\nu^{(T_h)}_h(s)\big\|_\infty, \big\|\log\bar\nu^{(T_h)}_h(s)\big\|_\infty\big\}\Big) \\
&\le 8H,
\end{aligned}$$

where the last step results from Lemma 24 (cf. (D.12)). Therefore, with $T_{\mathsf{start}} = \lceil \frac{1}{\eta\tau}\log H\rceil$ we can guarantee that

$$\begin{aligned}
\big\|Q^{(t)}_{h-1} - Q^\star_{h-1,\tau}\big\| &\le 10(1-\eta\tau)^{t-T_{h-1}}(1-\eta\tau)^{T_{\mathsf{start}}-1}\Big[H + \eta\tau t^{H-h+1}\Big] \\
&\le (1-\eta\tau)^{t-T_{h-1}}t^{H-h+1}.
\end{aligned}$$

This completes the proof for (5.16a). Regarding (5.16b), we start by the following lemmas, which are simply Lemma 21 and Lemma 23 applied to the episodic setting.

Lemma 21A. *With $0 < \eta \le \frac{1}{8H}$, we have*

$$\begin{aligned}
&\frac{1}{2}\mathsf{KL}_s\big(\zeta^\star_{h,\tau}\,\|\,\bar\zeta^{(t+1)}_h\big) + \eta\tau\mathsf{KL}_s\big(\bar\zeta^{(t+1)}_h\,\|\,\zeta^\star_{h,\tau}\big) \\
&\le (1-\eta\tau)\mathsf{KL}_s\big(\zeta^\star_{h,\tau}\,\|\,\zeta^{(t)}_h\big) + 2\eta H\mathsf{KL}_s\big(\zeta^{(t)}_h\,\|\,\bar\zeta^{(t)}_h\big) + 2\eta\big\|Q^{(t)}_h(s) - Q^\star_{h,\tau}(s)\big\|_\infty.
\end{aligned}$$

Lemma 23A. *It holds for all $h \in [H]$, $s \in \mathcal{S}$ and policy pair μ, ν that*

$$\max_{\mu',\nu'} \big(f_s(Q^\star_{h,\tau}, \mu'_h, \nu_h) - f_s(Q^\star_\tau, \mu_h, \nu'_h)\big) \le \frac{4H^2}{\tau}\mathsf{KL}_s\big(\zeta^\star_{h,\tau} \,\|\, \zeta_h\big) + \tau \mathsf{KL}_s\big(\zeta_h \,\|\, \zeta^\star_{h,\tau}\big).$$

We conclude that for $0 \le t_1 \le t_2 - 1$,

$$\begin{aligned}
&\max_{\mu,\nu} \big(f_s(Q^\star_{h,\tau}, \mu_h, \bar{\nu}_h^{(t_2)}) - f_s(Q^\star_\tau, \bar{\mu}_h^{(t_2)}, \nu_h)\big) \\
&\overset{\text{(i)}}{\le} \frac{4H^2}{\tau}\mathsf{KL}_s\big(\zeta^\star_{h,\tau} \,\|\, \bar{\zeta}_h^{(t_2)}\big) + \tau \mathsf{KL}_s\big(\bar{\zeta}_h^{(t_2)} \,\|\, \zeta^\star_{h,\tau}\big) \\
&\le \max\Big\{\frac{8H^2}{\tau}, \frac{1}{\eta}\Big\}\Big(\frac{1}{2}\mathsf{KL}_s\big(\zeta^\star_{h,\tau} \,\|\, \bar{\zeta}_h^{(t_2)}\big) + \eta\tau \mathsf{KL}_s\big(\bar{\zeta}_h^{(t_2)} \,\|\, \zeta^\star_{h,\tau}\big)\Big) \\
&\overset{\text{(ii)}}{\le} \max\Big\{\frac{8H^2}{\tau}, \frac{1}{\eta}\Big\}\Big((1-\eta\tau)\mathsf{KL}_s\big(\zeta^\star_{h,\tau} \,\|\, \zeta_h^{(t_2-1)}\big) + 2\eta H \mathsf{KL}_s\big(\zeta_h^{(t_2-1)} \,\|\, \bar{\zeta}_h^{(t_2-1)}\big) + 2\eta \big\|Q_h^{(t_2-1)}(s) - Q^\star_{h,\tau}(s)\big\|_\infty\Big) \\
&\overset{\text{(iii)}}{\le} \max\Big\{\frac{8H^2}{\tau}, \frac{1}{\eta}\Big\}\Big((1-\eta\tau)^{t_2-t_1}\Big(\mathsf{KL}_s\big(\zeta^\star_{h,\tau} \,\|\, \zeta_h^{(t_1)}\big) + (1-4\eta H)\mathsf{KL}_s\big(\zeta_h^{(t_1)} \,\|\, \bar{\zeta}_h^{(t_1)}\big)\Big) \\
&\qquad + 6\eta \sum_{l=t_1}^{t_2} (1-\eta\tau)^{t_2-l}\big\|Q_h^{(l)}(s) - Q^\star_{h,\tau}(s)\big\|_\infty\Big),
\end{aligned}$$

where (i) invokes Lemma 23A, (ii) invokes Lemma 21A and (iii) results from Lemma 25. It is straightforward to verify that the above inequality holds for $0 \le t_1 \le t_2$, by omitting the third step. Substitution of (5.16a) into the above inequality yields

$$\begin{aligned}
&\max_{\mu,\nu} \big(f_s(Q^\star_{h,\tau}, \mu_h, \bar{\nu}_h^{(t)}) - f_s(Q^\star_\tau, \bar{\mu}_h^{(t)}, \nu_h)\big) \\
&\le \max\Big\{\frac{8H^2}{\tau}, \frac{1}{\eta}\Big\}\Big((1-\eta\tau)^{t-T_h}\Big(\mathsf{KL}_s\big(\zeta^\star_{h,\tau} \,\|\, \zeta_h^{(T_h)}\big) + (1-4\eta H)\mathsf{KL}_s\big(\zeta_h^{(T_h)} \,\|\, \bar{\zeta}_h^{(T_h)}\big)\Big) \\
&\qquad + 6\eta \sum_{l=T_h}^{t} (1-\eta\tau)^{t-l}(1-\eta\tau)^{l-T_h} l^{H-h}\Big) \\
&\le (1-\eta\tau)^{t-T_h} \max\Big\{\frac{8H^2}{\tau}, \frac{1}{\eta}\Big\}\Big(\frac{8H}{\tau} + 6\eta t^{H-h+1}\Big).
\end{aligned} \tag{D.13}$$

We prove the following results instead, where (5.16b) is a direct consequence of (D.14) by summing up the two inequalities,

$$\begin{cases}
\max\limits_{s\in\mathcal{S},\mu} \Big(V_{h,\tau}^{\mu,\bar{\nu}^{(t)}}(s) - V^\star_{h,\tau}(s)\Big) \le 2(1-\eta\tau)^{t-T_h} \max\Big\{\frac{8H^2}{\tau}, \frac{1}{\eta}\Big\}\Big(\frac{8H}{\tau} + 6\eta t^{H-h+1}\Big) \\
\max\limits_{s\in\mathcal{S},\mu} \Big(V^\star_{h,\tau}(s) - V_{h,\tau}^{\bar{\mu}^{(t)},\nu}(s)\Big) \le 2(1-\eta\tau)^{t-T_h} \max\Big\{\frac{8H^2}{\tau}, \frac{1}{\eta}\Big\}\Big(\frac{8H}{\tau} + 6\eta t^{H-h+1}\Big)
\end{cases}. \tag{D.14}$$

We prove by induction. Note that when $h = H$, we have $V_{H,\tau}^{\mu,\nu}(s) = f_s(r_H, \mu_H, \nu_H) = f_s(Q^\star_{H,\tau}, \mu_H, \nu_H)$

and the claim holds by invoking (D.13). When the claim holds for some $2 \le h \le H$, we have

$$
\begin{aligned}
&V_{h-1,\tau}^{\mu,\bar{\nu}^{(t)}}(s) - V_{h-1,\tau}^{\star}(s) \\
&= \mu_{h-1}(s)^\top Q_{h-1,\tau}^{\mu,\bar{\nu}^{(t)}}(s)\bar{\nu}_{h-1}^{(t)}(s) + \tau\mathcal{H}\big(\mu_{h-1}(s)\big) - \tau\mathcal{H}\big(\bar{\nu}_{h-1}^{(t)}(s)\big) \\
&\quad - \mu_{h-1,\tau}^{\star}(s)^\top Q_{h-1,\tau}^{\star}(s)\nu_{h-1,\tau}^{\star}(s) + \tau\mathcal{H}\big(\mu_{h-1,\tau}^{\star}(s)\big) - \tau\mathcal{H}\big(\nu_{h-1,\tau}^{\star}(s)\big) \\
&= f_s(Q_{h-1,\tau}^{\star}, \mu_{h-1}, \bar{\nu}_{h-1}^{(t)}) - f_s(Q_{h-1,\tau}^{\star}, \mu_{h-1,\tau}^{\star}, \nu_{h-1,\tau}^{\star}) + \mu_{h-1}(s)^\top\big(Q_{h-1,\tau}^{\mu,\bar{\nu}^{(t)}}(s) - Q_{h-1,\tau}^{\star}(s)\big)\bar{\nu}_{h-1}^{(t)}(s) \\
&\le f_s(Q_{h-1,\tau}^{\star}, \mu_{h-1}, \bar{\nu}_{h-1}^{(t)}) - f_s(Q_{h-1,\tau}^{\star}, \bar{\mu}_{h-1}^{(t)}, \nu_{h-1,\tau}^{\star}) + \max_{s'\in\mathcal{S}}\Big[V_{h,\tau}^{\mu,\bar{\nu}^{(t)}}(s') - V_{h,\tau}^{\star}(s')\Big] \\
&\le \max_{\mu'_{h-1},\nu'_{h-1}}\Big(f_s(Q_{h-1,\tau}^{\star}, \mu'_{h-1}, \bar{\nu}_{h-1}^{(t)}) - f_s(Q_{h-1,\tau}^{\star}, \bar{\mu}_{h-1}^{(t)}, \nu'_{h-1})\Big) + \max_{s'\in\mathcal{S}}\Big[V_{h,\tau}^{\mu,\bar{\nu}^{(t)}}(s') - V_{h,\tau}^{\star}(s')\Big] \\
&\le (1-\eta\tau)^{t-T_{h-1}}\max\Big\{\frac{8H^2}{\tau}, \frac{1}{\eta}\Big\}\Big(\frac{8H}{\tau} + 6\eta t^{H-h+2}\Big) \\
&\qquad + 2(1-\eta\tau)^{t-T_h}\max\Big\{\frac{8H^2}{\tau}, \frac{1}{\eta}\Big\}\Big(\frac{8H}{\tau} + 6\eta t^{H-h+1}\Big) \\
&\le 2(1-\eta\tau)^{t-T_{h-1}}\max\Big\{\frac{8H^2}{\tau}, \frac{1}{\eta}\Big\}\Big(\frac{8H}{\tau} + 6\eta t^{H-h+2}\Big).
\end{aligned}
$$

Taking maximum over μ verifies the claim for $h-1$, thereby finishing the proof. The bound for $\max_{s\in\mathcal{S},\mu}\Big(V_{h,\tau}^{\star}(s) - V_{h,\tau}^{\bar{\mu}^{(t)},\nu}(s)\Big)$ can be established by following a similar argument and is therefore omitted.

D.3 Proof of key lemmas for the infinite-horizon setting

D.3.1 Proof of Lemma 17

Before proceeding, we shall introduce the following useful lemma that quantifies the distance between two consecutive updates, whose proof can be found in Appendix D.5.1.

Lemma 27. *For $0 < \eta \le 1/\tau$, it holds for all $s \in \mathcal{S}$ and $t \ge 0$ that*

$$
\max\Big\{\big\|\bar{\mu}^{(t+1)}(s) - \mu^{(t+1)}(s)\big\|_1, \big\|\bar{\nu}^{(t+1)}(s) - \nu^{(t+1)}(s)\big\|_1\Big\} \le \frac{2\eta}{1-\gamma}, \tag{D.15a}
$$

$$
\max\Big\{\big\|\bar{\mu}^{(t+1)}(s) - \bar{\mu}^{(t)}(s)\big\|_1, \big\|\bar{\nu}^{(t+1)}(s) - \bar{\nu}^{(t)}(s)\big\|_1\Big\} \le \frac{6\eta}{1-\gamma}, \tag{D.15b}
$$

and that

$$
\max\Big\{\big\|\log\zeta^{(t)}(s)\big\|_\infty, \big\|\log\bar{\zeta}^{(t)}(s)\big\|_\infty, \big\|\log\zeta_\tau^{\star}(s)\big\|_\infty\Big\} \le \frac{2}{(1-\gamma)\tau}. \tag{D.16}
$$

For notational simplicity, we use $x \overset{\mathbf{1}}{=} y$ to denote equivalence up to a global shift for two vectors x, y, i.e.

$$
x = y + c\cdot\mathbf{1} \tag{D.17}
$$

for some constant $c \in \mathbb{R}$. Taking logarithm on the both sides of the update rule (5.11a), we get

$$
\begin{cases}
\log\mu^{(t+1)}(s) - (1-\eta\tau)\log\mu^{(t)}(s) & \overset{\mathbf{1}}{=} \eta Q^{(t+1)}(s)\bar{\nu}^{(t+1)}(s) \\
\log\nu^{(t+1)}(s) - (1-\eta\tau)\log\nu^{(t)}(s) & \overset{\mathbf{1}}{=} -\eta Q^{(t+1)}(s)^\top\bar{\mu}^{(t+1)}(s)
\end{cases}. \tag{D.18}
$$

On the other hand, it holds for the QRE $(\mu_\tau^\star, \nu_\tau^\star)$ that

$$\begin{cases} \eta\tau \log \mu_\tau^\star(s) & \overset{\mathbf{1}}{=} \eta Q_\tau^\star(s)\nu_\tau^\star(s) \\ \eta\tau \log \nu_\tau^\star(s) & \overset{\mathbf{1}}{=} -\eta Q_\tau^\star(s)^\top \mu_\tau^\star(s) \end{cases}. \tag{D.19}$$

Subtracting (D.19) from (D.18) and taking inner product with $\bar{\zeta}^{(t+1)}(s) - \zeta_\tau^\star(s)$ gives

$$\begin{aligned}
&\left\langle \log \zeta^{(t+1)}(s) - (1-\eta\tau)\log \zeta^{(t)}(s) - \eta\tau \log \zeta_\tau^\star(s), \bar{\zeta}^{(t+1)}(s) - \zeta_\tau^\star(s) \right\rangle \\
&= \eta \left\langle \bar{\mu}^{(t+1)}(s) - \mu_\tau^\star(s), Q^{(t+1)}(s)\bar{\nu}^{(t+1)}(s) - Q_\tau^\star(s)\nu_\tau^\star(s) \right\rangle \\
&\quad - \eta \left\langle \bar{\nu}^{(t+1)}(s) - \nu_\tau^\star(s), Q^{(t+1)}(s)^\top \bar{\mu}^{(t+1)}(s) - Q_\tau^\star(s)^\top \mu_\tau^\star(s) \right\rangle \\
&= \eta \left\langle \bar{\mu}^{(t+1)}(s) - \mu_\tau^\star(s), (Q^{(t+1)}(s) - Q_\tau^\star(s))\bar{\nu}^{(t+1)}(s) \right\rangle \\
&\quad - \eta \left\langle \bar{\nu}^{(t+1)}(s) - \nu_\tau^\star(s), (Q^{(t+1)}(s) - Q_\tau^\star(s))^\top \bar{\mu}^{(t+1)}(s) \right\rangle \\
&= -\eta \left\langle \mu_\tau^\star(s), (Q^{(t+1)}(s) - Q_\tau^\star(s))\bar{\nu}^{(t+1)}(s) \right\rangle + \eta \left\langle \nu_\tau^\star(s), (Q^{(t+1)}(s) - Q_\tau^\star(s))^\top \bar{\mu}^{(t+1)}(s) \right\rangle \\
&\le 2\eta \left\| Q^{(t+1)}(s) - Q_\tau^\star(s) \right\|_\infty.
\end{aligned} \tag{D.20}$$

We continue to rewrite the LHS of (D.20) as

$$\begin{aligned}
&\left\langle \log \zeta^{(t+1)}(s) - (1-\eta\tau)\log \zeta^{(t)}(s) - \eta\tau \log \zeta_\tau^\star(s), \bar{\zeta}^{(t+1)}(s) - \zeta_\tau^\star(s) \right\rangle \\
&= -\left\langle \log \zeta^{(t+1)}(s) - (1-\eta\tau)\log \zeta^{(t)}(s) - \eta\tau \log \zeta_\tau^\star(s), \zeta_\tau^\star(s) \right\rangle \\
&\quad + \left\langle \log \bar{\zeta}^{(t+1)}(s) - (1-\eta\tau)\log \bar{\zeta}^{(t)}(s) - \eta\tau \log \zeta_\tau^\star(s), \bar{\zeta}^{(t+1)}(s) \right\rangle \\
&\quad + \left\langle \log \zeta^{(t+1)}(s) - \log \bar{\zeta}^{(t+1)}(s), \bar{\zeta}^{(t+1)}(s) \right\rangle \\
&\quad - (1-\eta\tau) \left\langle \log \zeta^{(t)}(s) - \log \bar{\zeta}^{(t)}(s), \bar{\zeta}^{(t+1)}(s) \right\rangle \\
&= \mathsf{KL}_s\big(\zeta_\tau^\star \,\|\, \zeta^{(t+1)}\big) - (1-\eta\tau)\mathsf{KL}_s\big(\zeta_\tau^\star \,\|\, \zeta^{(t)}\big) \\
&\quad + (1-\eta\tau)\mathsf{KL}_s\big(\bar{\zeta}^{(t+1)} \,\|\, \bar{\zeta}^{(t)}\big) + \eta\tau\mathsf{KL}_s\big(\bar{\zeta}^{(t+1)} \,\|\, \zeta_\tau^\star\big) \\
&\quad + \mathsf{KL}_s\big(\zeta^{(t+1)} \,\|\, \bar{\zeta}^{(t+1)}\big) - \left\langle \log \bar{\zeta}^{(t+1)}(s) - \log \zeta^{(t+1)}(s), \bar{\zeta}^{(t+1)}(s) - \zeta^{(t+1)}(s) \right\rangle \\
&\quad + (1-\eta\tau)\mathsf{KL}_s\big(\bar{\zeta}^{(t)} \,\|\, \zeta^{(t)}\big) - (1-\eta\tau)\left\langle \log \zeta^{(t)}(s) - \log \bar{\zeta}^{(t)}(s), \bar{\zeta}^{(t+1)}(s) - \bar{\zeta}^{(t)}(s) \right\rangle.
\end{aligned}$$

Rearranging terms, we have

$$\begin{aligned}
&\mathsf{KL}_s\big(\zeta_\tau^\star \,\|\, \zeta^{(t+1)}\big) - (1-\eta\tau)\mathsf{KL}_s\big(\zeta_\tau^\star \,\|\, \zeta^{(t)}\big) + (1-\eta\tau)\mathsf{KL}_s\big(\bar{\zeta}^{(t+1)} \,\|\, \bar{\zeta}^{(t)}\big) \\
&\quad + \eta\tau\mathsf{KL}_s\big(\bar{\zeta}^{(t+1)} \,\|\, \zeta_\tau^\star\big) + \mathsf{KL}_s\big(\zeta^{(t+1)} \,\|\, \bar{\zeta}^{(t+1)}\big) + (1-\eta\tau)\mathsf{KL}_s\big(\bar{\zeta}^{(t)} \,\|\, \zeta^{(t)}\big) \\
&\quad - \left\langle \log \bar{\zeta}^{(t+1)}(s) - \log \zeta^{(t+1)}(s), \bar{\zeta}^{(t+1)}(s) - \zeta^{(t+1)}(s) \right\rangle \\
&\quad - (1-\eta\tau)\left\langle \log \zeta^{(t)}(s) - \log \bar{\zeta}^{(t)}(s), \bar{\zeta}^{(t+1)}(s) - \bar{\zeta}^{(t)}(s) \right\rangle \\
&\le 2\eta \left\| Q^{(t+1)}(s) - Q_\tau^\star(s) \right\|_\infty.
\end{aligned}$$

It remains to upper bound

$$\left\langle \log \bar{\zeta}^{(t+1)}(s) - \log \zeta^{(t+1)}(s), \bar{\zeta}^{(t+1)}(s) - \zeta^{(t+1)}(s) \right\rangle \quad \text{and} \quad \left\langle \log \zeta^{(t)}(s) - \log \bar{\zeta}^{(t)}(s), \bar{\zeta}^{(t+1)}(s) - \bar{\zeta}^{(t)}(s) \right\rangle.$$

For the first term, note that

$$\begin{aligned}
&\Big\langle \log \bar{\mu}^{(t+1)}(s) - \log \mu^{(t+1)}(s), \bar{\mu}^{(t+1)}(s) - \mu^{(t+1)}(s) \Big\rangle \\
&= \eta \Big\langle Q^{(t)}(s)\bar{\nu}^{(t)}(s) - Q^{(t+1)}(s)\bar{\nu}^{(t+1)}(s), \bar{\mu}^{(t+1)}(s) - \mu^{(t+1)}(s) \Big\rangle \\
&\le \eta \big\| Q^{(t)}(s)\bar{\nu}^{(t)}(s) - Q^{(t+1)}(s)\bar{\nu}^{(t+1)}(s) \big\|_1 \big\| \bar{\mu}^{(t+1)}(s) - \mu^{(t+1)}(s) \big\|_1. \qquad \text{(D.21)}
\end{aligned}$$

Here, $\big\| Q^{(t)}(s)\bar{\nu}^{(t)}(s) - Q^{(t+1)}(s)\bar{\nu}^{(t+1)}(s) \big\|_1$ can be bounded as

$$\begin{aligned}
&\big\| Q^{(t)}(s)\bar{\nu}^{(t)}(s) - Q^{(t+1)}(s)\bar{\nu}^{(t+1)}(s) \big\|_1 \\
&\le \big\| Q^{(t+1)}(s)\big(\bar{\nu}^{(t)}(s) - \bar{\nu}^{(t+1)}(s)\big) \big\|_1 + \big\| \big(Q^{(t)}(s) - Q^{(t+1)}(s)\big)\bar{\nu}^{(t)}(s) \big\|_1 \\
&\le \frac{2}{1-\gamma} \big\| \bar{\nu}^{(t)}(s) - \bar{\nu}^{(t+1)}(s) \big\|_1 + \big\| Q^{(t)}(s) - Q^{(t+1)}(s) \big\|_\infty.
\end{aligned}$$

Plugging the above inequality into (D.21) and invoking Young's inequality yields

$$\begin{aligned}
&\Big\langle \log \bar{\mu}^{(t+1)}(s) - \log \mu^{(t+1)}(s), \bar{\mu}^{(t+1)}(s) - \mu^{(t+1)}(s) \Big\rangle \\
&\le \frac{\eta}{1-\gamma} \Big(\big\| \bar{\nu}^{(t+1)}(s) - \bar{\nu}^{(t)}(s) \big\|_1^2 + \big\| \bar{\mu}^{(t+1)}(s) - \mu^{(t+1)}(s) \big\|_1^2 \Big) \\
&\quad + \eta \big\| Q^{(t)}(s) - Q^{(t+1)}(s) \big\|_\infty \big\| \bar{\mu}^{(t+1)}(s) - \mu^{(t+1)}(s) \big\|_1 \\
&\le \frac{2\eta}{1-\gamma} \mathsf{KL}_s\big(\bar{\nu}^{(t+1)} \,\|\, \bar{\nu}^{(t)}\big) + \frac{2\eta}{1-\gamma} \mathsf{KL}_s\big(\mu^{(t+1)} \,\|\, \bar{\mu}^{(t+1)}\big) + \frac{2\eta^2}{1-\gamma} \big\| Q^{(t)}(s) - Q^{(t+1)}(s) \big\|_\infty, \qquad \text{(D.22)}
\end{aligned}$$

where the last step results from Pinsker's inequality and Lemma 27. Similarly, we have

$$\begin{aligned}
&\Big\langle \log \bar{\nu}^{(t+1)}(s) - \log \nu^{(t+1)}(s), \bar{\nu}^{(t+1)}(s) - \nu^{(t+1)}(s) \Big\rangle \\
&\le \frac{2\eta}{1-\gamma} \mathsf{KL}_s\big(\bar{\mu}^{(t+1)} \,\|\, \bar{\mu}^{(t)}\big) + \frac{2\eta}{1-\gamma} \mathsf{KL}_s\big(\nu^{(t+1)} \,\|\, \bar{\nu}^{(t+1)}\big) + \frac{2\eta^2}{1-\gamma} \big\| Q^{(t)}(s) - Q^{(t+1)}(s) \big\|_\infty.
\end{aligned}$$

Combining the above two inequalities gives

$$\begin{aligned}
&\Big\langle \log \bar{\zeta}^{(t+1)}(s) - \log \zeta^{(t+1)}(s), \bar{\zeta}^{(t+1)}(s) - \zeta^{(t+1)}(s) \Big\rangle \\
&\le \frac{2\eta}{1-\gamma} \mathsf{KL}_s\big(\bar{\zeta}^{(t+1)} \,\|\, \bar{\zeta}^{(t)}\big) + \frac{2\eta}{1-\gamma} \mathsf{KL}_s\big(\zeta^{(t+1)} \,\|\, \bar{\zeta}^{(t+1)}\big) + \frac{4\eta^2}{1-\gamma} \big\| Q^{(t)}(s) - Q^{(t+1)}(s) \big\|_\infty.
\end{aligned}$$

By a similar argument, when $t \ge 1$:

$$\begin{aligned}
&\Big\langle \log \zeta^{(t)}(s) - \log \bar{\zeta}^{(t)}(s), \bar{\zeta}^{(t+1)}(s) - \bar{\zeta}^{(t)}(s) \Big\rangle \\
&= \eta \Big\langle Q^{(t)}(s)\bar{\nu}^{(t)}(s) - Q^{(t-1)}(s)\bar{\nu}^{(t-1)}(s), \bar{\mu}^{(t+1)}(s) - \bar{\mu}^{(t)}(s) \Big\rangle \\
&\quad - \eta \Big\langle Q^{(t)}(s)^\top \bar{\mu}^{(t)}(s) - Q^{(t-1)}(s)^\top \bar{\mu}^{(t-1)}(s), \bar{\nu}^{(t+1)}(s) - \bar{\nu}^{(t)}(s) \Big\rangle \\
&\le \frac{2\eta}{1-\gamma} \mathsf{KL}_s\big(\bar{\zeta}^{(t)} \,\|\, \bar{\zeta}^{(t-1)}\big) + \frac{2\eta}{1-\gamma} \mathsf{KL}_s\big(\bar{\zeta}^{(t+1)} \,\|\, \bar{\zeta}^{(t)}\big) \\
&\quad + \eta \big(\big\| \bar{\mu}^{(t+1)}(s) - \bar{\mu}^{(t)}(s) \big\|_1 + \big\| \bar{\nu}^{(t+1)}(s) - \bar{\nu}^{(t)}(s) \big\|_1 \big) \big\| Q^{(t)}(s) - Q^{(t-1)}(s) \big\|_\infty \\
&\le \frac{2\eta}{1-\gamma} \mathsf{KL}_s\big(\bar{\zeta}^{(t)} \,\|\, \bar{\zeta}^{(t-1)}\big) + \frac{2\eta}{1-\gamma} \mathsf{KL}_s\big(\bar{\zeta}^{(t+1)} \,\|\, \bar{\zeta}^{(t)}\big) + \frac{12\eta^2}{1-\gamma} \big\| Q^{(t)}(s) - Q^{(t-1)}(s) \big\|_\infty.
\end{aligned}$$

Note that the above inequality trivially holds for $t = 0$, since $\log \zeta^{(0)}(s) = \log \bar{\zeta}^{(0)}(s)$.

Putting pieces together, we conclude that

$$\begin{aligned}
&\mathsf{KL}_s\big(\zeta_\tau^\star \,\|\, \zeta^{(t+1)}\big) - (1-\eta\tau)\mathsf{KL}_s\big(\zeta_\tau^\star \,\|\, \zeta^{(t)}\big) + \Big(1 - \eta\tau - \frac{4\eta}{1-\gamma}\Big)\mathsf{KL}_s\big(\bar{\zeta}^{(t+1)} \,\|\, \bar{\zeta}^{(t)}\big) \\
&\quad + \eta\tau\mathsf{KL}_s\big(\bar{\zeta}^{(t+1)} \,\|\, \zeta_\tau^\star\big) + \Big(1 - \frac{2\eta}{1-\gamma}\Big)\mathsf{KL}_s\big(\zeta^{(t+1)} \,\|\, \bar{\zeta}^{(t+1)}\big) + (1-\eta\tau)\mathsf{KL}_s\big(\bar{\zeta}^{(t)} \,\|\, \zeta^{(t)}\big) \\
&\quad - \frac{2\eta}{1-\gamma}\mathsf{KL}_s\big(\bar{\zeta}^{(t)} \,\|\, \bar{\zeta}^{(t-1)}\big) \\
&\le 2\eta\big\|Q^{(t+1)}(s) - Q_\tau^\star(s)\big\|_\infty + \frac{4\eta^2}{1-\gamma}\big\|Q^{(t)}(s) - Q^{(t+1)}(s)\big\|_\infty + \frac{12\eta^2}{1-\gamma}\big\|Q^{(t-1)}(s) - Q^{(t)}(s)\big\|_\infty.
\end{aligned}$$

Averaging state s over the initial state distribution ρ completes the proof.

D.3.2 Proof of Lemma 18

By definition of Q, it holds for $t \ge 1$ that

$$\big|Q^{(t+1)}(s,a,b) - Q^{(t)}(s,a,b)\big| \le \gamma\mathbb{E}_{s'\sim P(\cdot|s,a,b)}\Big[\big|V^{(t)}(s') - V^{(t-1)}(s')\big|\Big]. \tag{D.23}$$

Recall the definition of $f_s(Q,\mu,\nu)$ in (D.10) as the one-step entropy-regularized game value at state s, i.e.,

$$f_s(Q,\mu,\nu) = \mu(s)^\top Q(s)\nu(s) + \tau\mathcal{H}(\mu(s)) - \tau\mathcal{H}(\nu(s)),$$

which we further simplify the notation by introducing

$$f_s^{(t)} = f_s(Q^{(t)}, \bar{\mu}^{(t)}, \bar{\nu}^{(t)}).$$

By recursively applying the update rule $V^{(t)}(s) = (1-\alpha_t)V^{(t-1)}(s) + \alpha_t f_s^{(t)}$, we get

$$V^{(t)}(s) = \alpha_{0,t}V^{(0)} + \sum_{l=1}^{t}\alpha_{l,t}f_s(Q^{(l)}, \bar{\mu}^{(l)}, \bar{\nu}^{(l)}) = \sum_{l=0}^{t}\alpha_{l,t}f_s^{(l)}.$$

Therefore,

$$\begin{aligned}
\big|V^{(t)}(s) - V^{(t-1)}(s)\big| &= \alpha_t\big|f_s^{(t)} - V^{(t-1)}(s)\big| \\
&= \alpha_t\sum_{l=0}^{t-1}\alpha_{l,t-1}\big|f_s^{(t)} - f_s^{(l)}\big| \\
&\le \alpha_t\sum_{l=0}^{t-1}\alpha_{l,t-1}\sum_{j=l}^{t-1}\big|f_s^{(j+1)} - f_s^{(j)}\big|.
\end{aligned} \tag{D.24}$$

The next lemma enables us to upper bound $\big|f_s^{(t+1)} - f_s^{(t)}\big|$ with $\big\|Q^{(t+1)}(s) - Q^{(t)}(s)\big\|_\infty$ and $\mathsf{KL}_s\big(\bar{\zeta}^{(t+1)} \,\|\, \bar{\zeta}^{(t)}\big)$ as well as their counterparts in the $(t-1)$-th iteration. The proof is postponed to Appendix D.5.2.

Lemma 28. *For any $t \ge 0$, $\eta \le (1-\gamma)/180$, we have*

$$\begin{aligned}
\big|f_s^{(t+1)} - f_s^{(t)}\big| \le{}& \Big\|Q^{(t+1)}(s) - Q^{(t)}(s)\Big\|_\infty + \Big(\frac{3}{\eta} + \frac{4}{1-\gamma}\Big)\mathsf{KL}_s\big(\bar{\zeta}^{(t+1)} \,\|\, \bar{\zeta}^{(t)}\big) \\
&+ \frac{12\eta}{1-\gamma}\big\|Q^{(t)}(s) - Q^{(t-1)}(s)\big\|_\infty + \frac{2}{1-\gamma}\mathsf{KL}_s\big(\bar{\zeta}^{(t)} \,\|\, \bar{\zeta}^{(t-1)}\big).
\end{aligned}$$

Plugging the above lemma into (D.24),

$$
\begin{aligned}
&\big|V^{(t)}(s) - V^{(t-1)}(s)\big| \\
&\le \alpha_t \sum_{l=0}^{t-1} \alpha_{l,t-1} \sum_{j=l}^{t-1} \bigg[\big\|Q^{(j+1)}(s) - Q^{(j)}(s)\big\|_\infty + \Big(\frac{3}{\eta} + \frac{4}{1-\gamma}\Big) \mathsf{KL}_s\big(\bar{\zeta}^{(j+1)} \,\|\, \bar{\zeta}^{(j)}\big) \bigg] \\
&\quad + \alpha_t \sum_{l=0}^{t-1} \alpha_{l,t-1} \sum_{j=l}^{t-1} \bigg[\frac{12\eta}{1-\gamma} \big\|Q^{(j)}(s) - Q^{(j-1)}(s)\big\|_\infty + \frac{2}{1-\gamma} \mathsf{KL}_s\big(\bar{\zeta}^{(j)} \,\|\, \bar{\zeta}^{(j-1)}\big) \bigg] \\
&\le \alpha_t \sum_{l=0}^{t-1} \alpha_{l,t-1} \sum_{j=l}^{t-1} \bigg[\Big(1 + \frac{12\eta}{1-\gamma}\Big) \big\|Q^{(j+1)}(s) - Q^{(j)}(s)\big\|_\infty + \Big(\frac{3}{\eta} + \frac{6}{1-\gamma}\Big) \mathsf{KL}_s\big(\bar{\zeta}^{(j+1)} \,\|\, \bar{\zeta}^{(j)}\big) \bigg] \\
&\quad + \alpha_t \sum_{l=0}^{t-1} \alpha_{l,t-1} \bigg[\frac{12\eta}{1-\gamma} \big\|Q^{(l)}(s) - Q^{(l-1)}(s)\big\|_\infty + \frac{2}{1-\gamma} \mathsf{KL}_s\big(\bar{\zeta}^{(l)} \,\|\, \bar{\zeta}^{(l-1)}\big) \bigg] \\
&\le \sum_{j=0}^{t-1} \alpha_{j+1} \sum_{l=0}^{j} \alpha_{l,t-1} \bigg[\Big(1 + \frac{12\eta}{1-\gamma}\Big) \big\|Q^{(j+1)}(s) - Q^{(j)}(s)\big\|_\infty + \Big(\frac{3}{\eta} + \frac{6}{1-\gamma}\Big) \mathsf{KL}_s\big(\bar{\zeta}^{(j+1)} \,\|\, \bar{\zeta}^{(j)}\big) \bigg] \\
&\quad + \alpha_t \sum_{l=0}^{t-2} \alpha_{l+1,t-1} \bigg[\frac{12\eta}{1-\gamma} \big\|Q^{(l+1)}(s) - Q^{(l)}(s)\big\|_\infty + \frac{2}{1-\gamma} \mathsf{KL}_s\big(\bar{\zeta}^{(l+1)} \,\|\, \bar{\zeta}^{(l)}\big) \bigg],
\end{aligned}
$$

where the last step is due to $\alpha_t \le \alpha_j$ for all $j \le t$. To continue, by definition of α_t we have $\alpha_t \alpha_{l+1,t-1} \le \alpha_{l+1,t-1}(1-\alpha_t) = \alpha_{l+1,t}$ for $0 \le l < t$, and that

$$
\begin{aligned}
\alpha_{j+1} \sum_{l=0}^{j} \alpha_{l,t-1} &= \alpha_{j+1} \sum_{l=0}^{j} \Big(\prod_{i=l+1}^{t-1} (1-\alpha_i) - \prod_{i=l}^{t-1} (1-\alpha_i) \Big) \\
&= \alpha_{j+1} \prod_{i=j+1}^{t-1} (1-\alpha_i) \\
&\le \alpha_{j+1} \prod_{i=j+2}^{t} (1-\alpha_i) = \alpha_{j+1,t}.
\end{aligned}
$$

Plugging the inequality above into the previous relation gives

$$
\begin{aligned}
&\big|V^{(t)}(s) - V^{(t-1)}(s)\big| \\
&\le \sum_{j=0}^{t-1} \alpha_{j+1,t} \bigg[\Big(1 + \frac{12\eta}{1-\gamma}\Big) \big\|Q^{(j+1)}(s) - Q^{(j)}(s)\big\|_\infty + \Big(\frac{3}{\eta} + \frac{6}{1-\gamma}\Big) \mathsf{KL}_s\big(\bar{\zeta}^{(j+1)} \,\|\, \bar{\zeta}^{(j)}\big) \bigg] \\
&\quad + \sum_{l=0}^{t-2} \alpha_{l+1,t} \bigg[\frac{12\eta}{1-\gamma} \big\|Q^{(l+1)}(s) - Q^{(l)}(s)\big\|_\infty + \frac{2}{1-\gamma} \mathsf{KL}_s\big(\bar{\zeta}^{(l+1)} \,\|\, \bar{\zeta}^{(l)}\big) \bigg] \\
&\le \sum_{l=0}^{t-1} \alpha_{l+1,t} \bigg[\Big(1 + \frac{24\eta}{1-\gamma}\Big) \big\|Q^{(l+1)}(s) - Q^{(l)}(s)\big\|_\infty + \frac{4}{\eta} \mathsf{KL}_s\big(\bar{\zeta}^{(l+1)} \,\|\, \bar{\zeta}^{(l)}\big) \bigg].
\end{aligned}
$$

Plugging the above inequality into (D.23) leads to

$$\begin{aligned}
&\big|Q^{(t+1)}(s,a,b)-Q^{(t)}(s,a,b)\big| \\
&\le \gamma \mathop{\mathbb{E}}_{s'\sim P(\cdot|s,a,b))}\left\{\sum_{l=0}^{t-1}\alpha_{l+1,t}\Big[\Big(1+\frac{24\eta}{1-\gamma}\Big)\big\|Q^{(l+1)}(s')-Q^{(l)}(s')\big\|_\infty+\frac{4}{\eta}\mathsf{KL}_{s'}(\bar\zeta^{(l+1)}\,\|\,\bar\zeta^{(l)})\Big]\right\}.
\end{aligned}$$

When $\eta\le\frac{(1-\gamma)^2}{48\gamma}$, we have $\gamma(1+\frac{24\eta}{1-\gamma})\le\frac{1+\gamma}{2}$ and hence that

$$\begin{aligned}
&\big|Q^{(t+1)}(s,a,b)-Q^{(t)}(s,a,b)\big| \\
&\le \mathop{\mathbb{E}}_{s'\sim P(\cdot|s,a,b))}\left\{\frac{1+\gamma}{2}\sum_{l=0}^{t-1}\alpha_{l+1,t}\Big[\big\|Q^{(l+1)}(s')-Q^{(l)}(s')\big\|_\infty+\frac{4}{\eta}\mathsf{KL}_{s'}(\bar\zeta^{(l+1)}\,\|\,\bar\zeta^{(l)})\Big]\right\}.
\end{aligned}$$

Let $x^{(t+1)}\in\mathcal{A}^{\mathcal{S}}$ and $y^{(t+1)}\in\mathcal{B}^{\mathcal{S}}$ be defined as for any $s\in\mathcal{S}$:

$$(x^{(t+1)}(s),y^{(t+1)}(s))=\arg\max_{(a,b)\in\mathcal{A}\times\mathcal{B}}\big|Q^{(t+1)}(s,a,b)-Q^{(t)}(s,a,b)\big|.$$

It follows that $\forall\chi\in\Gamma(\rho)$, we have $\chi P_{x^{(t+1)},y^{(t+1)}}\in\Gamma(\rho)$ and hence

$$\begin{aligned}
&\mathop{\mathbb{E}}_{s\sim\chi}\Big[\big\|Q^{(t+1)}(s)-Q^{(t)}(s)\big\|_\infty\Big] \\
&=\mathop{\mathbb{E}}_{\substack{s\sim\chi,\\ a=x^{(t+1)}(s),b=y^{(t+1)}(s)}}\Big[\big|Q^{(t+1)}(s,a,b)-Q^{(t)}(s,a,b)\big|\Big] \\
&\le \mathop{\mathbb{E}}_{s'\sim\chi P_{x^{(t+1)},y^{(t+1)}}}\left[\frac{1+\gamma}{2}\sum_{l=0}^{t-1}\alpha_{l+1,t}\Big[\big\|Q^{(l+1)}(s')-Q^{(l)}(s')\big\|_\infty+\frac{4}{\eta}\mathsf{KL}_{s'}(\bar\zeta^{(l+1)}\,\|\,\bar\zeta^{(l)})\Big]\right] \\
&\le \frac{1+\gamma}{2}\sum_{l=0}^{t-1}\alpha_{l+1,t}\left[\big\|Q^{(l+1)}-Q^{(l)}\big\|_{\Gamma(\rho)}+\frac{4}{\eta}\cdot\Big\|\frac{\chi P_{x^{(t+1)},y^{(t+1)}}}{\rho}\Big\|_\infty\mathsf{KL}_\rho\big(\bar\zeta^{(l+1)}\,\|\,\bar\zeta^{(l)}\big)\right] \\
&\le \frac{1+\gamma}{2}\sum_{l=0}^{t-1}\alpha_{l+1,t}\left[\big\|Q^{(l+1)}-Q^{(l)}\big\|_{\Gamma(\rho)}+\frac{4\mathcal{C}_\rho}{\eta}\mathsf{KL}_\rho\big(\bar\zeta^{(l+1)}\,\|\,\bar\zeta^{(l)}\big)\right]. \qquad \text{(D.25)}
\end{aligned}$$

Taking the supremum over $\chi\in\Gamma(\rho)$ completes the proof for $t\ge1$. To complete the proof, note that when $t=0$, we have $\big\|Q^{(0)}-Q^{(1)}\big\|_{\Gamma(\rho)}=\big\|Q^{(1)}\big\|_{\Gamma(\rho)}\le2$.

D.3.3 Proof of Lemma 19

Note that it suffices to show for $t\ge0$, $s\in\mathcal{S}$, $(a,b)\in\mathcal{A}\times\mathcal{B}$:

$$\begin{aligned}
&\big|Q^{(t+1)}(s,a,b)-Q_\tau^\star(s,a,b)\big| \\
&\le\frac{1+\gamma}{2}\cdot\mathop{\mathbb{E}}_{s'\sim P(s,a,b)}\left[\sum_{l=0}^{t}\alpha_{l,t}\Big[\big\|Q^{(l)}(s')-Q_\tau^\star(s')\big\|_\infty+\frac{2\eta}{1-\gamma}\big\|Q^{(l)}(s')-Q^{(l-1)}(s')\big\|_\infty\Big]\right]+2\alpha_{0,t}. \qquad \text{(D.26)}
\end{aligned}$$

The remaining step follows a similar argument as (D.25) and is therefore omitted.

To establish (D.26), notice that we have for $t \geq 0$,

$$Q^{(t+1)}(s,a,b) - Q_\tau^\star(s,a,b) = \gamma \mathbb{E}_{s' \sim P(\cdot|s,a,b)} \left[V^{(t)}(s') - V_\tau^\star(s') \right] = \gamma \mathbb{E}_{s' \sim P(\cdot|s,a,b)} \left[\sum_{l=0}^{t} \alpha_{l,t} (f_{s'}^{(l)} - f_{s'}^\star) \right]. \tag{D.27}$$

To continue, we start by decomposing $f_s^{(t)} - f_s^\star$ as

$$\begin{aligned}
f_s^{(t)} - f_s^\star &= f_s(Q^{(t)}, \bar{\mu}^{(t)}, \bar{\nu}^{(t)}) - f_s(Q_\tau^\star, \mu_\tau^\star, \nu_\tau^\star) \\
&= \Big(f_s(Q^{(t)}, \bar{\mu}^{(t)}, \bar{\nu}^{(t)}) - f_s(Q^{(t)}, \bar{\mu}^{(t)}, \nu_\tau^\star) \Big) + f_s(Q^{(t)}, \bar{\mu}^{(t)}, \nu_\tau^\star) - f_s(Q_\tau^\star, \mu_\tau^\star, \nu_\tau^\star) \\
&\leq \Big(f_s(Q^{(t)}, \bar{\mu}^{(t)}, \bar{\nu}^{(t)}) - f_s(Q^{(t)}, \bar{\mu}^{(t)}, \nu_\tau^\star) \Big) + f_s(Q_\tau^\star, \bar{\mu}^{(t)}, \nu_\tau^\star) - f_s(Q_\tau^\star, \mu_\tau^\star, \nu_\tau^\star) \\
&\quad + \big\| Q^{(t)}(s) - Q_\tau^\star(s) \big\|_\infty \\
&\leq f_s(Q^{(t)}, \bar{\mu}^{(t)}, \bar{\nu}^{(t)}) - f_s(Q^{(t)}, \bar{\mu}^{(t)}, \nu_\tau^\star) + \big\| Q^{(t)}(s) - Q_\tau^\star(s) \big\|_\infty.
\end{aligned}$$

We bound the first two terms with the following lemma, whose proof can be found in Appendix D.5.3.

Lemma 29. *It holds for all $t \geq 0$, $s \in \mathcal{S}$ and $\nu(s) \in \Delta(\mathcal{B})$ that*

$$\begin{aligned}
&f_s(Q^{(t)}, \bar{\mu}^{(t)}, \bar{\nu}^{(t)}) - f_s(Q^{(t)}, \bar{\mu}^{(t)}, \nu) \\
&\leq \frac{2\eta}{1-\gamma} \big\| Q^{(t)}(s) - Q^{(t-1)}(s) \big\|_\infty + \frac{2}{1-\gamma} \Big(\mathsf{KL}_s\big(\bar{\mu}^{(t)} \,\|\, \mu^{(t-1)}\big) + \mathsf{KL}_s\big(\mu^{(t-1)} \,\|\, \bar{\mu}^{(t-1)}\big) \Big) \\
&\quad - \frac{1}{\eta}\big(1 - \frac{4\eta}{1-\gamma}\big) \mathsf{KL}_s\big(\nu^{(t)} \,\|\, \bar{\nu}^{(t)}\big) - \frac{1-\eta\tau}{\eta} \mathsf{KL}_s\big(\bar{\nu}^{(t)} \,\|\, \nu^{(t-1)}\big) \\
&\quad + \frac{1-\eta\tau}{\eta} \mathsf{KL}_s\big(\nu \,\|\, \nu^{(t-1)}\big) - \frac{1}{\eta} \mathsf{KL}_s\big(\nu \,\|\, \nu^{(t)}\big).
\end{aligned}$$

Applying Lemma 29 with $\nu(s) = \nu_\tau^\star(s)$ gives

$$\begin{aligned}
f_s^{(t)} - f_s^\star \leq{}& \big\| Q^{(t)}(s) - Q_\tau^\star(s) \big\|_\infty + \frac{2\eta}{1-\gamma} \big\| Q^{(t)}(s) - Q^{(t-1)}(s) \big\|_\infty \\
&+ \frac{1-\eta\tau}{\eta} \mathsf{KL}_s\big(\nu_\tau^\star \,\|\, \nu^{(t-1)}\big) - \frac{1}{\eta} \mathsf{KL}_s\big(\nu_\tau^\star \,\|\, \nu^{(t)}\big) \\
&- \frac{1}{\eta}\big(1 - \frac{4\eta}{1-\gamma}\big) \mathsf{KL}_s\big(\nu^{(t)} \,\|\, \bar{\nu}^{(t)}\big) - \frac{1-\eta\tau}{\eta} \mathsf{KL}_s\big(\bar{\nu}^{(t)} \,\|\, \nu^{(t-1)}\big) \\
&+ \frac{2}{1-\gamma} \Big(\mathsf{KL}_s\big(\bar{\mu}^{(t)} \,\|\, \mu^{(t-1)}\big) + \mathsf{KL}_s\big(\mu^{(t-1)} \,\|\, \bar{\mu}^{(t-1)}\big) \Big).
\end{aligned} \tag{D.28}$$

By a similar argument, we can derive

$$\begin{aligned}
f_s^\star - f_s^{(t)} \leq{}& \big\| Q^{(t)}(s) - Q_\tau^\star(s) \big\|_\infty + \frac{2\eta}{1-\gamma} \big\| Q^{(t)}(s) - Q^{(t-1)}(s) \big\|_\infty \\
&+ \frac{1-\eta\tau}{\eta} \mathsf{KL}_s\big(\mu_\tau^\star \,\|\, \mu^{(t-1)}\big) - \frac{1}{\eta} \mathsf{KL}_s\big(\mu_\tau^\star \,\|\, \mu^{(t)}\big) \\
&- \frac{1}{\eta}\big(1 - \frac{4\eta}{1-\gamma}\big) \mathsf{KL}_s\big(\mu^{(t)} \,\|\, \bar{\mu}^{(t)}\big) - \frac{1-\eta\tau}{\eta} \mathsf{KL}_s\big(\bar{\mu}^{(t)} \,\|\, \mu^{(t-1)}\big) \\
&+ \frac{2}{1-\gamma} \Big(\mathsf{KL}_s\big(\bar{\nu}^{(t)} \,\|\, \nu^{(t-1)}\big) + \mathsf{KL}_s\big(\nu^{(t-1)} \,\|\, \bar{\nu}^{(t-1)}\big) \Big).
\end{aligned} \tag{D.29}$$

Combining (D.28) $+\frac{1-\gamma}{4}\cdot$ (D.29) gives

$$
\begin{aligned}
&(1-\frac{1-\gamma}{4})(f_s^{(t)}-f_s^\star)\\
&\le (1+\frac{1-\gamma}{4})\Big[\big\|Q^{(t)}(s)-Q_\tau^\star(s)\big\|_\infty+\frac{2\eta}{1-\gamma}\big\|Q^{(t)}(s)-Q^{(t-1)}(s)\big\|_\infty\Big]\\
&\quad+\frac{1-\eta\tau}{\eta}\Big[\mathsf{KL}_s\big(\nu_\tau^\star\,\|\,\nu^{(t-1)}\big)+\frac{1-\gamma}{4}\mathsf{KL}_s\big(\mu_\tau^\star\,\|\,\mu^{(t-1)}\big)\Big]-\frac{1}{\eta}\Big[\mathsf{KL}_s\big(\nu_\tau^\star\,\|\,\nu^{(t)}\big)+\frac{1-\gamma}{4}\mathsf{KL}_s\big(\mu_\tau^\star\,\|\,\mu^{(t)}\big)\Big]\\
&\quad+\frac{2}{1-\gamma}\Big[\mathsf{KL}_s\big(\mu^{(t-1)}\,\|\,\bar\mu^{(t-1)}\big)+\frac{1-\gamma}{4}\mathsf{KL}_s\big(\nu^{(t-1)}\,\|\,\bar\nu^{(t-1)}\big)\Big]\\
&\quad-\frac{1}{\eta}\big(1-\frac{4\eta}{1-\gamma}\big)\Big[\frac{1-\gamma}{4}\mathsf{KL}_s\big(\mu^{(t)}\,\|\,\bar\mu^{(t)}\big)+\mathsf{KL}_s\big(\nu^{(t)}\,\|\,\bar\nu^{(t)}\big)\Big]\\
&\quad+(\frac{2}{1-\gamma}-\frac{1-\eta\tau}{\eta}\cdot\frac{1-\gamma}{4})\mathsf{KL}_s\big(\bar\mu^{(t)}\,\|\,\mu^{(t-1)}\big)+(\frac{2}{1-\gamma}\cdot\frac{1-\gamma}{4}-\frac{1-\eta\tau}{\eta})\mathsf{KL}_s\big(\bar\nu^{(t)}\,\|\,\nu^{(t-1)}\big).
\end{aligned}
\tag{D.30}
$$

With $0<\eta\le(1-\gamma)^2/16$, we have

$$
\begin{aligned}
&\frac{2}{1-\gamma}-\frac{1-\eta\tau}{\eta}\cdot\frac{1-\gamma}{4})\le 0,\qquad \frac{2}{1-\gamma}\cdot\frac{1-\gamma}{4}-\frac{1-\eta\tau}{\eta}\le 0,\\
&\frac{1}{\eta}\big(1-\frac{4\eta}{1-\gamma}\big)\cdot\frac{1-\gamma}{4}\ge\frac{2}{1-\gamma}\cdot\frac{1}{1-\eta\tau}.
\end{aligned}
$$

To proceed, we introduce a shorthand notation

$$
\begin{aligned}
G^{(t)}(s)=&\frac{1}{\eta}\Big[\mathsf{KL}_s\big(\nu_\tau^\star\,\|\,\nu^{(t)}\big)+\frac{1-\gamma}{4}\mathsf{KL}_s\big(\mu_\tau^\star\,\|\,\mu^{(t)}\big)\Big]\\
&+\frac{2}{(1-\gamma)(1-\eta\tau)}\Big[\mathsf{KL}_s\big(\mu^{(t)}\,\|\,\bar\mu^{(t)}\big)+\mathsf{KL}_s\big(\nu^{(t)}\,\|\,\bar\nu^{(t)}\big)\Big].
\end{aligned}
$$

We can then write (D.30) as

$$
\begin{aligned}
(1-\frac{1-\gamma}{4})(f_s^{(t)}-f_s^\star)\le&(1+\frac{1-\gamma}{4})\Big[\big\|Q^{(t)}(s)-Q_\tau^\star(s)\big\|_\infty+\frac{2\eta}{1-\gamma}\big\|Q^{(t)}(s)-Q^{(t-1)}(s)\big\|_\infty\Big]\\
&+(1-\eta\tau)G^{(t-1)}(s)-G^{(t)}(s).
\end{aligned}
\tag{D.31}
$$

Note that when $t=0$, we have

$$
\begin{aligned}
f_s^{(0)}-f_s^\star&=\tau\log|\mathcal{A}|-\tau\log|\mathcal{B}|-\mu_\tau^\star(s)^\top Q_\tau^\star(s)\nu_\tau^\star(s)-\tau\mathcal{H}(\mu_\tau^\star(s))+\tau\mathcal{H}(\nu_\tau^\star(s))\\
&=\max_{\mu(s)}\min_{\nu(s)}f_s(Q^{(0)},\mu,\nu)-\max_{\mu(s)}\min_{\nu(s)}f_s(Q_\tau^\star,\mu,\nu)\\
&\le\big\|Q^{(0)}(s)-Q_\tau^\star(s)\big\|_\infty.
\end{aligned}
\tag{D.32}
$$

Substitution of (D.31) and (D.32) into (D.27) gives

$$
\begin{aligned}
&Q^{(t+1)}(s,a,b) - Q_\tau^\star(s,a,b) \\
&= \gamma \mathbb{E}_{s'\sim P(\cdot|s,a,b)} \left[\sum_{l=0}^{t} \alpha_{l,t} (f_{s'}^{(l)} - f_{s'}^\star) \right] \\
&\le \gamma \mathbb{E}_{s'\sim P(s,a,b)} \left[\alpha_{0,t} \left\| Q^{(0)}(s') - Q_\tau^\star(s') \right\|_\infty \right] \\
&\quad + \gamma \cdot \frac{1+(1-\gamma)/4}{1-(1-\gamma)/4} \mathop{\mathbb{E}}_{s'\sim P(s,a,b)} \left[\sum_{l=1}^{t} \alpha_{l,t} \left[\left\| Q^{(l)}(s') - Q_\tau^\star(s') \right\|_\infty + \frac{2\eta}{1-\gamma} \left\| Q^{(l)}(s') - Q^{(l-1)}(s') \right\|_\infty \right] \right] \\
&\quad + \frac{\gamma}{1-(1-\gamma)/4} \mathop{\mathbb{E}}_{s'\sim P(s,a,b)} \left[(1-\eta\tau) \sum_{l=1}^{t} \alpha_{l,t} G^{(l-1)}(s') - \sum_{l=1}^{t} \alpha_{l,t} G^{(l)}(s') \right].
\end{aligned}
$$

Note that

$$
\begin{aligned}
(1-\eta\tau) \sum_{l=1}^{t} \alpha_{l,t} G^{(l-1)}(s') - \sum_{l=1}^{t} \alpha_{l,t} G^{(l)}(s') &\le \sum_{l=1}^{t-1} ((1-\eta\tau)\alpha_{l+1,t} - \alpha_{l,t}) G^{(l)}(s') + \alpha_{1,t} G^{(0)}(s') \\
&\le \alpha_{1,t} G^{(0)}(s') \le 2\alpha_{0,t} \eta\tau G^{(0)}(s') \le 2\alpha_{0,t},
\end{aligned}
$$

where the second step is due to

$$
\begin{aligned}
(1-\eta\tau)\alpha_{l+1,t} - \alpha_{l,t} &= ((1-\eta\tau)\alpha_{l+1} - \alpha_l(1-\alpha_{l+1})) \prod_{j=l+2}^{t} \alpha_j \\
&\le ((1-\eta\tau)\alpha_{l+1} - \alpha_{l+1} + \alpha_l\alpha_{l+1}) \prod_{j=l+2}^{t} \alpha_j \\
&= \alpha_{l+1}(\alpha_l - \eta\tau) \prod_{j=l+2}^{t} \alpha_j \le 0. \qquad \text{(D.33)}
\end{aligned}
$$

We conclude that

$$
\begin{aligned}
&Q^{(t+1)}(s,a,b) - Q_\tau^\star(s,a,b) \\
&\le \gamma \cdot \frac{1+(1-\gamma)/4}{1-(1-\gamma)/4} \mathop{\mathbb{E}}_{s'\sim P(s,a,b)} \left[\sum_{l=0}^{t} \alpha_{l,t} \left[\left\| Q^{(l)}(s') - Q_\tau^\star(s') \right\|_\infty + \frac{2\eta}{1-\gamma} \left\| Q^{(l)}(s') - Q^{(l-1)}(s') \right\|_\infty \right] \right] \\
&\quad + 2\alpha_{0,t} \\
&\le \frac{1+\gamma}{2} \cdot \mathop{\mathbb{E}}_{s'\sim P(s,a,b)} \left[\sum_{l=0}^{t} \alpha_{l,t} \left[\left\| Q^{(l)}(s') - Q_\tau^\star(s') \right\|_\infty + \frac{2\eta}{1-\gamma} \left\| Q^{(l)}(s') - Q^{(l-1)}(s') \right\|_\infty \right] \right] \\
&\quad + 2\alpha_{0,t}.
\end{aligned}
$$

The other side of (D.26) can be obtained by computing $\frac{1-\gamma}{4}\cdot$ (D.28) + (D.29) and following a similar argument, and is therefore omitted. To conclude the proof, we note that for $t=0$, we have $\left|Q^{(1)}(s,a,b) - Q_\tau^\star(s,a,b)\right| \le \gamma \max_{s'\in\mathcal{S}} |f_{s'}^{(0)} - f_{s'}^\star| \le \frac{2\gamma}{1-\gamma}$.

D.3.4 Proof of Lemma 20

For $t \geq 1$, let

$$u_t = \eta \big\| Q_\tau^\star(s) - Q^{(t)}(s) \big\|_{\Gamma(\rho)} + \frac{12\eta^2}{(1-\gamma)^2} \big\| Q^{(t)}(s) - Q^{(t-1)}(s) \big\|_{\Gamma(\rho)}.$$

It follows that

$$u_1 \leq \frac{2\gamma\eta}{1-\gamma} + \frac{24\eta^2}{(1-\gamma)^3} \leq 1.$$

When $t \geq 1$, invoking Lemma 18 and Lemma 19 gives

$$\begin{aligned} u_{t+1} &\leq \Big(1 - \frac{1-\gamma}{2}\Big) \sum_{l=1}^{t} \alpha_{l,t} \Big[\eta \big\| Q^{(l)} - Q_\tau^\star \big\|_{\Gamma(\rho)} + \Big(\frac{2\eta^2}{1-\gamma} + \frac{12\eta^2}{(1-\gamma)^2} \Big) \big\| Q^{(l)} - Q^{(l-1)} \big\|_{\Gamma(\rho)} \Big] \\ &\quad + \frac{48\eta\mathcal{C}_\rho}{(1-\gamma)^2} \sum_{l=1}^{t} \alpha_{l,t} \mathsf{KL}_\rho\big(\bar{\zeta}^{(l)} \,\|\, \bar{\zeta}^{(l-1)}\big) + 2\alpha_{0,t}\eta + \alpha_{0,t}\eta \big\| Q^{(0)} - Q_\tau^\star \big\|_{\Gamma(\rho)} \qquad \text{(D.34)} \\ &\leq \Big(1 - \frac{1-\gamma}{3}\Big) \sum_{l=1}^{t} \alpha_{l,t} u_l + \frac{48\eta\mathcal{C}_\rho}{(1-\gamma)^2} \sum_{l=1}^{t} \alpha_{l,t} \mathsf{KL}_\rho\big(\bar{\zeta}^{(l)} \,\|\, \bar{\zeta}^{(l-1)}\big) + \frac{4\eta}{1-\gamma}\alpha_{0,t}. \end{aligned}$$

Let

$$\beta_{l,t} = \alpha_l \prod_{i=l+1}^{t} \Big(1 - \frac{1-\gamma}{3} \cdot \alpha_i\Big).$$

It follows that for $t \geq 0$,

$$\begin{aligned} &\sum_{l=1}^{t+1} \alpha_{l,t+1} u_l \\ &= (1-\alpha_{t+1}) \sum_{l=1}^{t} \alpha_{l,t} u_l + \alpha_{t+1} u_{t+1} \\ &\leq \Big(1 - \frac{1-\gamma}{3} \cdot \alpha_{t+1}\Big) \sum_{l=1}^{t} \alpha_{l,t} u_l + \alpha_{t+1} \frac{48\eta\mathcal{C}_\rho}{(1-\gamma)^2} \cdot \sum_{l=1}^{t} \alpha_{l,t} \mathsf{KL}_\rho\big(\bar{\zeta}^{(l)} \,\|\, \bar{\zeta}^{(l-1)}\big) + \frac{4\eta}{1-\gamma}\alpha_{t+1}\alpha_{0,t} \\ &\leq \prod_{l=2}^{t+1} \Big(1 - \frac{1-\gamma}{3} \cdot \alpha_l\Big) \alpha_{1,1} u_1 + \frac{48\eta\mathcal{C}_\rho}{(1-\gamma)^2} \sum_{i=1}^{t} \beta_{i+1,t+1} \sum_{l=1}^{i} \alpha_{l,i} \mathsf{KL}_\rho\big(\bar{\zeta}^{(l)} \,\|\, \bar{\zeta}^{(l-1)}\big) + \frac{4\eta}{1-\gamma} \sum_{i=1}^{t} \alpha_{0,i}\beta_{i+1,t+1} \\ &\leq \beta_{1,t+1} u_1 + \frac{48\eta\mathcal{C}_\rho}{(1-\gamma)^2} \sum_{l=1}^{t} \sum_{i=l}^{t} \alpha_{l,i}\beta_{i+1,t+1} \mathsf{KL}_\rho\big(\bar{\zeta}^{(l)} \,\|\, \bar{\zeta}^{(l-1)}\big) + \frac{4\eta}{1-\gamma} \sum_{i=1}^{t} \alpha_{0,i}\beta_{i+1,t+1} \\ &\leq \frac{200\eta\mathcal{C}_\rho}{(1-\gamma)^2} \sum_{l=1}^{t} \beta_{l,t+1} \mathsf{KL}_\rho\big(\bar{\zeta}^{(l)} \,\|\, \bar{\zeta}^{(l-1)}\big) + \frac{18\eta}{1-\gamma}\beta_{0,t+1}, \qquad \text{(D.35)} \end{aligned}$$

where the last step is due to the following lemma. Similar lemma has appeared in prior works (see i.e., [Wei et al., 2021b, Lemma 36]). Our version features a simpler proof, which is postponed to Appendix D.5.4.

Lemma 30. *Let two sequences $\{\delta_i\}, \{\xi_i\}$ be defined as*

$$\delta_i = 1 - c_1\alpha_i, \qquad \text{and} \qquad \xi_i = 1 - c_2\alpha_i,$$

where the constants c_1, c_2 satisfy $0 < c_1 < c_2 < \frac{1}{2\alpha_i}$. For $l \le t$, let $\delta_{l,t} = \alpha_l \prod_{i=l+1}^{t} \delta_i$ and $\xi_{l,t} = \alpha_l \prod_{i=l+1}^{t} \xi_i$, where we take $\delta_{l,l} = \xi_{l,l} = \alpha_l$. We have

$$\sum_{i=l}^{t} \xi_{l,i}\delta_{i+1,t} \le \Big(1 + \frac{2}{c_2 - c_1}\Big)\delta_{l,t}.$$

Substitution of (D.35) into (D.34) gives

$$\begin{aligned}
u_{t+1} &\le \Big(1 - \frac{1-\gamma}{3}\Big)\sum_{l=1}^{t} \alpha_{l,t}u_l + \frac{48\eta}{(1-\gamma)^2}\sum_{l=1}^{t}\alpha_{l,t}\mathsf{KL}_\rho\big(\bar{\zeta}^{(l)} \,\|\, \bar{\zeta}^{(l-1)}\big) + \frac{4\eta}{1-\gamma}\alpha_{0,t} \\
&\le \frac{200\eta\mathcal{C}_\rho}{(1-\gamma)^2}\sum_{l=1}^{t}\beta_{l,t}\mathsf{KL}_\rho\big(\bar{\zeta}^{(l)} \,\|\, \bar{\zeta}^{(l-1)}\big) + \frac{18\eta}{1-\gamma}\beta_{0,t} + \frac{48\eta\mathcal{C}_\rho}{(1-\gamma)^2}\sum_{l=1}^{t}\alpha_{l,t}\mathsf{KL}_\rho\big(\bar{\zeta}^{(l)} \,\|\, \bar{\zeta}^{(l-1)}\big) + \frac{4\eta}{1-\gamma}\alpha_{0,t} \\
&\le \frac{250\eta\mathcal{C}_\rho}{(1-\gamma)^2}\sum_{l=1}^{t}\beta_{l,t}\mathsf{KL}_\rho\big(\bar{\zeta}^{(l)} \,\|\, \bar{\zeta}^{(l-1)}\big) + \frac{22\eta}{1-\gamma}\beta_{0,t}.
\end{aligned}$$

for $t \ge 1$. It is straightforward to verify that the above inequality holds for $t = 0$ as well.

So we conclude that

$$\begin{aligned}
&\sum_{l=0}^{t}\lambda_{l+1,t+1}u_{l+1} = \sum_{i=0}^{t}\lambda_{i+1,t+1}u_{i+1} \\
&\le \sum_{i=0}^{t}\lambda_{i+1,t+1}\Big[\frac{250\eta\mathcal{C}_\rho}{(1-\gamma)^2}\sum_{l=1}^{i}\beta_{l,i}\mathsf{KL}_\rho\big(\bar{\zeta}^{(l)} \,\|\, \bar{\zeta}^{(l-1)}\big) + \frac{22\eta}{1-\gamma}\beta_{0,i}\Big] \\
&= \frac{250\eta\mathcal{C}_\rho}{(1-\gamma)^2}\sum_{l=1}^{t}\sum_{i=l}^{t}\beta_{l,i}\lambda_{i+1,t+1}\mathsf{KL}_\rho\big(\bar{\zeta}^{(l)} \,\|\, \bar{\zeta}^{(l-1)}\big) + \frac{22\eta}{1-\gamma}\sum_{i=0}^{t}\beta_{0,i}\lambda_{i+1,t+1} \\
&\le \frac{6250\eta\mathcal{C}_\rho}{(1-\gamma)^3}\sum_{l=1}^{t}\lambda_{l,t+1}\mathsf{KL}_\rho\big(\bar{\zeta}^{(l)} \,\|\, \bar{\zeta}^{(l-1)}\big) + \frac{550\eta}{(1-\gamma)^2}\lambda_{0,t+1} \\
&= \frac{6250\eta\mathcal{C}_\rho}{(1-\gamma)^3}\sum_{l=0}^{t-1}\lambda_{l+1,t+1}\mathsf{KL}_\rho\big(\bar{\zeta}^{(l+1)} \,\|\, \bar{\zeta}^{(l)}\big) + \frac{550\eta}{(1-\gamma)^2}\lambda_{0,t+1},
\end{aligned}$$

where the penultimate step invokes Lemma 30.

D.3.5 Proof of Lemma 21

Taking logarithm on the both sides of the update rule (5.11b), we get

$$\begin{cases}
\log\bar{\mu}^{(t+1)}(s) - (1-\eta\tau)\log\mu^{(t)}(s) & \overset{\mathbf{1}}{=} \eta Q^{(t)}(s)\bar{\nu}^{(t)}(s) \\
\log\bar{\nu}^{(t+1)}(s) - (1-\eta\tau)\log\nu^{(t)}(s) & \overset{\mathbf{1}}{=} -\eta Q^{(t)}(s)^\top\bar{\mu}^{(t)}(s)
\end{cases}, \tag{D.36}$$

where we recall the notation in (D.17).

Subtracting (D.19) from (D.36) and taking inner product with $\bar{\zeta}^{(t+1)}(s) - \zeta_\tau^\star(s)$ gives

$$
\begin{aligned}
&\Big\langle \log \bar{\zeta}^{(t+1)}(s) - (1-\eta\tau)\log \zeta^{(t)}(s) - \eta\tau \log \zeta_\tau^\star(s), \bar{\zeta}^{(t+1)}(s) - \zeta_\tau^\star(s) \Big\rangle \\
&= \eta \Big\langle \bar{\mu}^{(t+1)}(s) - \mu_\tau^\star(s), Q^{(t)}(s)\bar{\nu}^{(t)}(s) - Q_\tau^\star(s)\nu_\tau^\star(s) \Big\rangle \\
&\qquad - \eta \Big\langle \bar{\nu}^{(t+1)}(s) - \nu_\tau^\star(s), Q^{(t)}(s)^\top \bar{\mu}^{(t)}(s) - Q_\tau^\star(s)^\top \mu_\tau^\star(s) \Big\rangle \\
&\le \eta \Big\langle \bar{\mu}^{(t+1)}(s) - \mu_\tau^\star(s), Q^{(t)}(s)\big(\bar{\nu}^{(t)}(s) - \nu_\tau^\star(s)\big) \Big\rangle \\
&\qquad - \eta \Big\langle \bar{\nu}^{(t+1)}(s) - \nu_\tau^\star(s), Q^{(t)}(s)^\top \big(\bar{\mu}^{(t)}(s) - \mu_\tau^\star(s)\big) \Big\rangle + 2\eta \big\| Q^{(t)}(s) - Q_\tau^\star(s) \big\|_\infty \\
&\le \eta \Big\langle \bar{\mu}^{(t+1)}(s) - \mu_\tau^\star(s), Q^{(t)}(s)\big(\bar{\nu}^{(t)}(s) - \bar{\nu}^{(t+1)}(s)\big) \Big\rangle \\
&\qquad - \eta \Big\langle \bar{\nu}^{(t+1)}(s) - \nu_\tau^\star(s), Q^{(t)}(s)^\top \big(\bar{\mu}^{(t)}(s) - \bar{\mu}^{(t+1)}(s)\big) \Big\rangle + 2\eta \big\| Q^{(t)}(s) - Q_\tau^\star(s) \big\|_\infty \\
&\le \frac{2\eta}{1-\gamma}\Big(2\mathsf{KL}_s\big(\zeta_\tau^\star \,\|\, \bar{\zeta}^{(t+1)}\big) + \mathsf{KL}_s\big(\bar{\zeta}^{(t+1)} \,\|\, \zeta^{(t)}\big) + \mathsf{KL}_s\big(\zeta^{(t)} \,\|\, \bar{\zeta}^{(t)}\big)\Big) + 2\eta \big\| Q^{(t)}(s) - Q_\tau^\star(s) \big\|_\infty .
\end{aligned}
$$

LHS can be written as

$$
\begin{aligned}
&\Big\langle \log \bar{\zeta}^{(t+1)}(s) - (1-\eta\tau)\log \zeta^{(t)}(s) - \eta\tau \log \zeta_\tau^\star(s), \bar{\zeta}^{(t+1)}(s) - \zeta_\tau^\star(s) \Big\rangle \\
&= -\Big\langle \log \bar{\zeta}^{(t+1)}(s) - (1-\eta\tau)\log \zeta^{(t)}(s) - \eta\tau \log \zeta_\tau^\star(s), \zeta_\tau^\star(s) \Big\rangle \\
&\qquad + \Big\langle \log \bar{\zeta}^{(t+1)}(s) - (1-\eta\tau)\log \zeta^{(t)}(s) - \eta\tau \log \zeta_\tau^\star(s), \bar{\zeta}^{(t+1)}(s) \Big\rangle \\
&= \mathsf{KL}_s\big(\zeta_\tau^\star \,\|\, \bar{\zeta}^{(t+1)}\big) - (1-\eta\tau)\mathsf{KL}_s\big(\zeta_\tau^\star \,\|\, \zeta^{(t)}\big) + (1-\eta\tau)\mathsf{KL}_s\big(\bar{\zeta}^{(t+1)} \,\|\, \zeta^{(t)}\big) + \eta\tau\mathsf{KL}_s\big(\bar{\zeta}^{(t+1)} \,\|\, \zeta_\tau^\star\big).
\end{aligned}
$$

So we conclude that

$$
\begin{aligned}
&\Big(1 - \frac{4\eta}{1-\gamma}\Big)\mathsf{KL}_s\big(\zeta_\tau^\star \,\|\, \bar{\zeta}^{(t+1)}\big) - (1-\eta\tau)\mathsf{KL}_s\big(\zeta_\tau^\star \,\|\, \zeta^{(t)}\big) \\
&\qquad + \Big(1 - \eta\tau - \frac{2\eta}{1-\gamma}\Big)\mathsf{KL}_s\big(\bar{\zeta}^{(t+1)} \,\|\, \zeta^{(t)}\big) + \eta\tau\mathsf{KL}_s\big(\bar{\zeta}^{(t+1)} \,\|\, \zeta_\tau^\star\big) \\
&\le \frac{2\eta}{1-\gamma}\mathsf{KL}_s\big(\zeta^{(t)} \,\|\, \bar{\zeta}^{(t)}\big) + 2\eta \big\| Q^{(t)}(s) - Q_\tau^\star(s) \big\|_\infty .
\end{aligned}
$$

With $0 < \eta \le \frac{1-\gamma}{8}$, we have

$$
\begin{aligned}
&\frac{1}{2}\mathsf{KL}_s\big(\zeta_\tau^\star \,\|\, \bar{\zeta}^{(t+1)}\big) + \eta\tau\mathsf{KL}_s\big(\bar{\zeta}^{(t+1)} \,\|\, \zeta_\tau^\star\big) \\
&\le (1-\eta\tau)\mathsf{KL}_s\big(\zeta_\tau^\star \,\|\, \zeta^{(t)}\big) + \frac{2\eta}{1-\gamma}\mathsf{KL}_s\big(\zeta^{(t)} \,\|\, \bar{\zeta}^{(t)}\big) + 2\eta \big\| Q^{(t)}(s) - Q_\tau^\star(s) \big\|_\infty .
\end{aligned}
$$

D.3.6 Proof of Lemma 22

By definition of value function V_τ, we have

$$
\begin{aligned}
V_\tau^{\mu,\nu}(s) - V_\tau^\star(s) &= \mu(s)^\top Q_\tau^{\mu,\nu}(s)\nu(s) + \tau\mathcal{H}\big(\mu(s)\big) - \tau\mathcal{H}\big(\nu(s)\big) \\
&\qquad - \mu_\tau^\star(s)^\top Q_\tau^\star(s)\nu_\tau^\star(s) - \tau\mathcal{H}\big(\mu_\tau^\star(s)\big) + \tau\mathcal{H}\big(\nu_\tau^\star(s)\big) \\
&= \mu(s)^\top Q_\tau^{\mu,\nu}(s)\nu(s) - \mu(s)^\top Q_\tau^\star(s)\nu(s) + f_s(Q_\tau^\star, \mu, \nu) - f_s(Q_\tau^\star, \mu_\tau^\star, \nu_\tau^\star) \\
&= \gamma \mathop{\mathbb{E}}_{\substack{a\sim\mu(\cdot|s), b\sim\nu(\cdot|s), \\ s'\sim P(\cdot|s,a,b)}} \big[V_\tau^{\mu,\nu}(s') - V_\tau^\star(s')\big] + f_s(Q_\tau^\star, \mu, \nu) - f_s(Q_\tau^\star, \mu_\tau^\star, \nu_\tau^\star).
\end{aligned}
$$

Applying the relation recursively and averaging s over ρ, we arrive at

$$V_\tau^{\mu,\nu}(\rho) - V_\tau^\star(\rho) = \frac{1}{1-\gamma} \mathop{\mathbb{E}}_{s' \sim d_\rho^{\mu,\nu}} \left[f_{s'}(Q_\tau^\star, \mu, \nu) - f_{s'}(Q_\tau^\star, \mu_\tau^\star, \nu_\tau^\star) \right], \tag{D.37}$$

which is the well-known performance difference lemma applied to the setting of Markov games. It follows that

$$\begin{aligned}
V_\tau^{\mu_\tau^\dagger(\nu),\nu}(\rho) - V_\tau^\star(\rho) &= \frac{1}{1-\gamma} \mathop{\mathbb{E}}_{s' \sim d_\rho^{\mu_\tau^\dagger(\nu),\nu}} \left[f_{s'}(Q_\tau^\star, \mu_\tau^\dagger(\nu), \nu) - f_{s'}(Q_\tau^\star, \mu_\tau^\star, \nu_\tau^\star) \right] \\
&\le \frac{1}{1-\gamma} \mathop{\mathbb{E}}_{s' \sim d_\rho^{\mu_\tau^\dagger(\nu),\nu}} \left[f_{s'}(Q_\tau^\star, \mu_\tau^\dagger(\nu), \nu) - f_{s'}(Q_\tau^\star, \mu, \nu_\tau^\star) \right] \\
&\le \frac{1}{1-\gamma} \mathop{\mathbb{E}}_{s' \sim d_\rho^{\mu_\tau^\dagger(\nu),\nu}} \left[\max_{\mu',\nu'} \left(f_{s'}(Q_\tau^\star, \mu', \nu) - f_{s'}(Q_\tau^\star, \mu, \nu') \right) \right] \\
&\le \frac{\mathcal{C}_{\rho,\tau}^\dagger}{1-\gamma} \mathop{\mathbb{E}}_{s \sim \rho} \left[\max_{\mu',\nu'} \left(f_s(Q_\tau^\star, \mu', \nu) - f_s(Q_\tau^\star, \mu, \nu') \right) \right].
\end{aligned} \tag{D.38}$$

A similar argument gives $V_\tau^\star(\rho) - V_\tau^{\mu,\nu_\tau^\dagger(\mu)}(\rho) \le \frac{\mathcal{C}_{\rho,\tau}^\dagger}{1-\gamma} \mathop{\mathbb{E}}_{s \sim \rho} \left[\max_{\mu',\nu'} \left(f_s(Q_\tau^\star, \mu', \nu) - f_s(Q_\tau^\star, \mu, \nu') \right) \right]$. Summing the two inequalities proves (D.8). Alternatively, we continue from (D.38) and show that

$$\begin{aligned}
V_\tau^{\mu_\tau^\dagger(\nu),\nu}(s) - V_\tau^\star(s) &\le \frac{1}{1-\gamma} \mathop{\mathbb{E}}_{s' \sim d_s^{\mu_\tau^\dagger(\nu),\nu}} \left[\max_{\mu',\nu'} \left(f_{s'}(Q_\tau^\star, \mu', \nu) - f_{s'}(Q_\tau^\star, \mu, \nu') \right) \right] \\
&\le \frac{\|1/\rho\|_\infty}{1-\gamma} \mathop{\mathbb{E}}_{s \sim \rho} \left[\max_{\mu',\nu'} \left(f_s(Q_\tau^\star, \mu', \nu) - f_s(Q_\tau^\star, \mu, \nu') \right) \right].
\end{aligned}$$

Summing the inequality with the one for $V_\tau^\star(s) - V_\tau^{\mu,\nu_\tau^\dagger(\mu)}(s)$ and taking maximum over $s \in \mathcal{S}$ completes the proof for (D.9).

D.4 Proof of key lemmas for the finite-horizon setting

D.4.1 Proof of Lemma 25

Following similar arguments of arriving (D.20), we have

$$\begin{aligned}
&\left\langle \log \zeta_h^{(t+1)}(s) - (1-\eta\tau) \log \zeta_h^{(t)}(s) - \eta\tau \log \zeta_{h,\tau}^\star(s), \bar{\zeta}_h^{(t+1)}(s) - \zeta_{h,\tau}^\star(s) \right\rangle \\
&\le 2\eta \left\| Q_h^{(t+1)}(s) - Q_{h,\tau}^\star(s) \right\|_\infty.
\end{aligned}$$

We rewrite the LHS as

$$
\begin{aligned}
&\Big\langle \log \zeta_h^{(t+1)}(s) - (1-\eta\tau)\log \zeta_h^{(t)}(s) - \eta\tau \log \zeta_{h,\tau}^{\star}(s), \bar{\zeta}_h^{(t+1)}(s) - \zeta_{h,\tau}^{\star}(s)\Big\rangle \\
&= -\Big\langle \log \zeta_h^{(t+1)}(s) - (1-\eta\tau)\log \zeta_h^{(t)}(s) - \eta\tau \log \zeta_{h,\tau}^{\star}(s), \zeta_{h,\tau}^{\star}(s)\Big\rangle \\
&\quad + \Big\langle \log \bar{\zeta}_h^{(t+1)}(s) - (1-\eta\tau)\log \zeta_h^{(t)}(s) - \eta\tau \log \zeta_{h,\tau}^{\star}(s), \bar{\zeta}_h^{(t+1)}(s)\Big\rangle \\
&\quad + \Big\langle \log \zeta_h^{(t+1)}(s) - \log \bar{\zeta}_h^{(t+1)}(s), \bar{\zeta}_h^{(t+1)}(s)\Big\rangle \\
&= \mathsf{KL}_s\big(\zeta_{h,\tau}^{\star} \,\|\, \zeta_h^{(t+1)}\big) - (1-\eta\tau)\mathsf{KL}_s\big(\zeta_{h,\tau}^{\star} \,\|\, \zeta_h^{(t)}\big) \\
&\quad + (1-\eta\tau)\mathsf{KL}_s\big(\bar{\zeta}_h^{(t+1)} \,\|\, \zeta_h^{(t)}\big) + \eta\tau\mathsf{KL}_s\big(\bar{\zeta}_h^{(t+1)} \,\|\, \zeta_{h,\tau}^{\star}\big) \\
&\quad + \mathsf{KL}_s\big(\zeta_h^{(t+1)} \,\|\, \bar{\zeta}_h^{(t+1)}\big) - \Big\langle \log \bar{\zeta}_h^{(t+1)}(s) - \log \zeta_h^{(t+1)}(s), \bar{\zeta}_h^{(t+1)}(s) - \zeta_h^{(t+1)}(s)\Big\rangle.
\end{aligned}
$$

Rearranging terms gives

$$
\begin{aligned}
&\mathsf{KL}_s\big(\zeta_{h,\tau}^{\star} \,\|\, \zeta_h^{(t+1)}\big) - (1-\eta\tau)\mathsf{KL}_s\big(\zeta_{h,\tau}^{\star} \,\|\, \zeta_h^{(t)}\big) + (1-\eta\tau)\mathsf{KL}_s\big(\bar{\zeta}_h^{(t+1)} \,\|\, \zeta_h^{(t)}\big) \\
&\qquad + \eta\tau\mathsf{KL}_s\big(\bar{\zeta}_h^{(t+1)} \,\|\, \zeta_{h,\tau}^{\star}\big) + \mathsf{KL}_s\big(\zeta_h^{(t+1)} \,\|\, \bar{\zeta}_h^{(t+1)}\big) \\
&\qquad - \Big\langle \log \bar{\zeta}_h^{(t+1)}(s) - \log \zeta_h^{(t+1)}(s), \bar{\zeta}_h^{(t+1)}(s) - \zeta_h^{(t+1)}(s)\Big\rangle \\
&\le 2\eta \big\| Q^{(t+1)}(s) - Q_{\tau}^{\star}(s)\big\|_{\infty}. \qquad \text{(D.39)}
\end{aligned}
$$

Note that

$$
\begin{aligned}
&\Big\langle \log \bar{\mu}_h^{(t+1)}(s) - \log \mu_h^{(t+1)}(s), \bar{\mu}_h^{(t+1)}(s) - \mu_h^{(t+1)}(s)\Big\rangle \\
&= \eta \Big\langle Q_h^{(t)}(s)\bar{\nu}_h^{(t)}(s) - Q_h^{(t+1)}(s)\bar{\nu}_h^{(t+1)}(s), \bar{\mu}_h^{(t+1)}(s) - \mu_h^{(t+1)}(s)\Big\rangle \\
&\le \eta \big\| Q_h^{(t)}(s)\bar{\nu}_h^{(t)}(s) - Q_h^{(t+1)}(s)\bar{\nu}_h^{(t+1)}(s)\big\|_1 \big\|\bar{\mu}_h^{(t+1)}(s) - \mu_h^{(t+1)}(s)\big\|_1. \qquad \text{(D.40)}
\end{aligned}
$$

We bound $\big\|Q_h^{(t)}(s)\bar{\nu}_h^{(t)}(s) - Q_h^{(t+1)}(s)\bar{\nu}_h^{(t+1)}(s)\big\|_1$ as

$$
\begin{aligned}
&\big\|Q_h^{(t)}(s)\bar{\nu}_h^{(t)}(s) - Q_h^{(t+1)}(s)\bar{\nu}_h^{(t+1)}(s)\big\|_1 \\
&\le \big\|Q_h^{(t+1)}(s)\big(\bar{\nu}_h^{(t)}(s) - \bar{\nu}_h^{(t+1)}(s)\big)\big\|_1 + \big\|\big(Q_h^{(t)}(s) - Q_h^{(t+1)}(s)\big)\bar{\nu}_h^{(t)}(s)\big\|_1 \\
&\le 2H\big\|\bar{\nu}_h^{(t)}(s) - \bar{\nu}_h^{(t+1)}(s)\big\|_1 + \big\|Q_h^{(t)}(s) - Q_h^{(t+1)}(s)\big\|_{\infty} \\
&\le 2H\big\|\bar{\nu}_h^{(t+1)}(s) - \nu_h^{(t)}(s)\big\|_1 + 2H\big\|\nu_h^{(t)}(s) - \bar{\nu}_h^{(t)}(s)\big\|_1 + \big\|Q_h^{(t)}(s) - Q_h^{(t+1)}(s)\big\|_{\infty}.
\end{aligned}
$$

Plugging the above inequality into (D.40) and invoking Young's inequality yields

$$
\begin{aligned}
&\Big\langle \log \bar{\mu}_h^{(t+1)}(s) - \log \mu_h^{(t+1)}(s), \bar{\mu}_h^{(t+1)}(s) - \mu_h^{(t+1)}(s)\Big\rangle \\
&\le \eta H\Big(\big\|\bar{\nu}_h^{(t+1)}(s) - \nu_h^{(t)}(s)\big\|_1^2 + \big\|\nu_h^{(t)}(s) - \bar{\nu}_h^{(t)}(s)\big\|_1^2 + 2\big\|\bar{\mu}_h^{(t+1)}(s) - \mu_h^{(t+1)}(s)\big\|_1^2\Big) \\
&\qquad + \eta\big\|Q_h^{(t)}(s) - Q_h^{(t+1)}(s)\big\|_{\infty}\big\|\bar{\mu}_h^{(t+1)}(s) - \mu_h^{(t+1)}(s)\big\|_1 \\
&\le 2\eta H\mathsf{KL}_s\big(\bar{\nu}_h^{(t+1)} \,\|\, \nu_h^{(t)}\big) + 2\eta H\mathsf{KL}_s\big(\nu_h^{(t)} \,\|\, \bar{\nu}_h^{(t)}\big) + 4\eta H\mathsf{KL}_s\big(\mu_h^{(t+1)} \,\|\, \bar{\mu}_h^{(t+1)}\big) + 2\eta^2 H\big\|Q_h^{(t)}(s) - Q_h^{(t+1)}(s)\big\|_{\infty},
\end{aligned}
$$

where the last step results from Pinsker's inequality and Lemma 24. Similarly, we have

$$
\begin{aligned}
&\Big\langle \log \bar{\nu}_h^{(t+1)}(s) - \log \nu_h^{(t+1)}(s), \bar{\nu}_h^{(t+1)}(s) - \nu_h^{(t+1)}(s)\Big\rangle \\
&\le 2\eta H\mathsf{KL}_s\big(\bar{\mu}_h^{(t+1)} \,\|\, \mu_h^{(t)}\big) + 2\eta H\mathsf{KL}_s\big(\mu_h^{(t)} \,\|\, \bar{\mu}_h^{(t)}\big) + 4\eta H\mathsf{KL}_s\big(\nu_h^{(t+1)} \,\|\, \bar{\nu}_h^{(t+1)}\big) + 2\eta^2 H\big\|Q_h^{(t)}(s) - Q_h^{(t+1)}(s)\big\|_{\infty}.
\end{aligned}
$$

Summing the above two inequalities gives

$$
\begin{aligned}
&\Big\langle \log \bar{\zeta}_h^{(t+1)}(s) - \log \zeta_h^{(t+1)}(s), \bar{\zeta}_h^{(t+1)}(s) - \zeta_h^{(t+1)}(s) \Big\rangle \\
&\le 2\eta H \mathsf{KL}_s\big(\bar{\zeta}_h^{(t+1)} \,\|\, \zeta_h^{(t)}\big) + 2\eta H \mathsf{KL}_s\big(\zeta_h^{(t)} \,\|\, \bar{\zeta}_h^{(t)}\big) + 4\eta H \mathsf{KL}_s\big(\zeta_h^{(t+1)} \,\|\, \bar{\zeta}_h^{(t+1)}\big) \\
&\qquad + 4\eta^2 H \big\|Q_h^{(t)}(s) - Q_h^{(t+1)}(s)\big\|_\infty \\
&\le 2\eta H \mathsf{KL}_s\big(\bar{\zeta}_h^{(t+1)} \,\|\, \zeta_h^{(t)}\big) + 2\eta H \mathsf{KL}_s\big(\zeta_h^{(t)} \,\|\, \bar{\zeta}_h^{(t)}\big) + 4\eta H \mathsf{KL}_s\big(\zeta_h^{(t+1)} \,\|\, \bar{\zeta}_h^{(t+1)}\big) \\
&\qquad + \frac{\eta}{2}\Big(\big\|Q_h^{(t)}(s) - Q_{h,\tau}^\star(s)\big\|_\infty + \big\|Q_h^{(t+1)}(s) - Q_{h,\tau}^\star(s)\big\|_\infty\Big),
\end{aligned}
$$

where the second step invokes triangular inequality and the fact that $\eta \le \frac{1}{8H}$. Plugging the above inequality into (D.39) gives

$$
\begin{aligned}
&\mathsf{KL}_s\big(\zeta_{h,\tau}^\star \,\|\, \zeta_h^{(t+1)}\big) - (1-\eta\tau)\mathsf{KL}_s\big(\zeta_{h,\tau}^\star \,\|\, \zeta_h^{(t)}\big) + (1-\eta(\tau+2H))\mathsf{KL}_s\big(\bar{\zeta}_h^{(t+1)} \,\|\, \zeta_h^{(t)}\big) \\
&\qquad + \eta\tau \mathsf{KL}_s\big(\bar{\zeta}_h^{(t+1)} \,\|\, \zeta_{h,\tau}^\star\big) + (1-4\eta H)\mathsf{KL}_s\big(\zeta_h^{(t+1)} \,\|\, \bar{\zeta}_h^{(t+1)}\big) - 2\eta H \mathsf{KL}_s\big(\zeta_h^{(t)} \,\|\, \bar{\zeta}_h^{(t)}\big) \\
&\le \frac{5\eta}{2}\big\|Q_h^{(t+1)}(s) - Q_\tau^\star(s)\big\|_\infty + \frac{\eta}{2}\big\|Q_h^{(t)}(s) - Q_\tau^\star(s)\big\|_\infty.
\end{aligned}
$$

With $\eta \le \frac{1}{8H}$, we have $(1-\eta\tau)(1-4\eta H) \ge 2\eta H$ and $1-\eta(\tau+2H) \ge 0$. It follows that

$$
\begin{aligned}
&\mathsf{KL}_s\big(\zeta_{h,\tau}^\star \,\|\, \zeta_h^{(t+1)}\big) + (1-4\eta H)\mathsf{KL}_s\big(\zeta_h^{(t+1)} \,\|\, \bar{\zeta}_h^{(t+1)}\big) + \eta\tau \mathsf{KL}_s\big(\bar{\zeta}_h^{(t+1)} \,\|\, \zeta_{h,\tau}^\star\big) \\
&\le (1-\eta\tau)\mathsf{KL}_s\big(\zeta_{h,\tau}^\star \,\|\, \zeta_h^{(t)}\big) + 2\eta H \mathsf{KL}_s\big(\zeta_h^{(t)} \,\|\, \bar{\zeta}_h^{(t)}\big) \\
&\qquad + \frac{5\eta}{2}\big\|Q_h^{(t+1)}(s) - Q_\tau^\star(s)\big\|_\infty + \frac{\eta}{2}\big\|Q_h^{(t)}(s) - Q_\tau^\star(s)\big\|_\infty \\
&\le (1-\eta\tau)\Big(\mathsf{KL}_s\big(\zeta_{h,\tau}^\star \,\|\, \zeta_h^{(t)}\big) + (1-4\eta H)\mathsf{KL}_s\big(\zeta_h^{(t)} \,\|\, \bar{\zeta}_h^{(t)}\big)\Big) \\
&\qquad + \frac{5\eta}{2}\big\|Q_h^{(t+1)}(s) - Q_\tau^\star(s)\big\|_\infty + \frac{\eta}{2}\big\|Q_h^{(t)}(s) - Q_\tau^\star(s)\big\|_\infty.
\end{aligned}
$$

Therefore, it holds for $0 \le t_1 < t_2$ that

$$
\begin{aligned}
&\mathsf{KL}_s\big(\zeta_{h,\tau}^\star \,\|\, \zeta_h^{(t_2)}\big) + (1-4\eta H)\mathsf{KL}_s\big(\zeta_h^{(t_2)} \,\|\, \bar{\zeta}_h^{(t_2)}\big) + \eta\tau \mathsf{KL}_s\big(\bar{\zeta}_h^{(t_2)} \,\|\, \zeta_{h,\tau}^\star\big) \\
&\le (1-\eta\tau)^{t_2-t_1}\Big(\mathsf{KL}_s\big(\zeta_{h,\tau}^\star \,\|\, \zeta_h^{(t_1)}\big) + (1-4\eta H)\mathsf{KL}_s\big(\zeta_h^{(t_1)} \,\|\, \bar{\zeta}_h^{t_1}\big)\Big) \\
&\qquad + \sum_{t'=t_1+1}^{t_2} (1-\eta\tau)^{t_2-l}\Big[\frac{5\eta}{2}\big\|Q_h^{(l)}(s) - Q_\tau^\star(s)\big\|_\infty + \frac{\eta}{2}\big\|Q_h^{(l-1)}(s) - Q_\tau^\star(s)\big\|_\infty\Big] \\
&\le (1-\eta\tau)^{t_2-t_1}\Big(\mathsf{KL}_s\big(\zeta_{h,\tau}^\star \,\|\, \zeta_h^{(t_1)}\big) + (1-4\eta H)\mathsf{KL}_s\big(\zeta_h^{(t_1)} \,\|\, \bar{\zeta}_h^{(t_1)}\big)\Big) \\
&\qquad + 4\eta \sum_{l=t_1}^{t_2} (1-\eta\tau)^{t_2-l}\big\|Q_h^{(l)}(s) - Q_\tau^\star(s)\big\|_\infty.
\end{aligned}
$$

D.4.2 Proof of Lemma 26

For $t_2 > 0$, we have

$$
\begin{aligned}
&Q_{h-1}^{(t_2)}(s,a,b) - Q_{h-1,\tau}^{\star}(s,a,b) \\
&= \mathop{\mathbb{E}}_{s'\sim P_{h-1}(\cdot|s,a,b)} \Big[V_h^{(t_2-1)}(s') - V_{h,\tau}^{\star}(s')\Big] \\
&= \mathop{\mathbb{E}}_{s'\sim P_{h-1}(\cdot|s,a,b)} \Big[(1-\eta\tau)^{t_2-t_1}\big(V_h^{(t_1-1)}(s') - V_{h,\tau}^{\star}(s')\big) \\
&\qquad + \eta\tau \sum_{l=t_1}^{t_2-1} (1-\eta\tau)^{t_2-1-l}\big(f_{s'}(Q^{(t_1)}, \bar{\mu}_h^{(t_1)}, \bar{\nu}_h^{(t_1)}) - f_{s'}(Q_{h,\tau}^{\star}, \mu_{h,\tau}^{\star}, \nu_{h,\tau}^{\star})\big)\Big] \\
&\le (1-\eta\tau)^{t_2-t_1} 2H + \mathop{\mathbb{E}}_{s'\sim P_{h-1}(\cdot|s,a,b)} \left[\eta\tau \sum_{l=t_1}^{t_2-1} (1-\eta\tau)^{t_2-1-l}\big(f_{s'}(Q_h^{(l)}, \bar{\mu}_h^{(l)}, \bar{\nu}_h^{(l)}) - f_{s'}(Q_{h,\tau}^{\star}, \mu_{h,\tau}^{\star}, \nu_{h,\tau}^{\star})\big)\right].
\end{aligned} \tag{D.41}
$$

We start by decomposing $f_s^{(t)} - f_s^{\star}$ as

$$
\begin{aligned}
&f_s(Q_h^{(t)}, \bar{\mu}_h^{(t)}, \bar{\nu}_h^{(t)}) - f_s(Q_{h,\tau}^{\star}, \mu_{h,\tau}^{\star}, \nu_{h,\tau}^{\star}) \\
&= \Big(f_s(Q_h^{(t)}, \bar{\mu}_h^{(t)}, \bar{\nu}_h^{(t)}) - f_s(Q_h^{(t)}, \bar{\mu}_h^{(t)}, \nu_{h,\tau}^{\star})\Big) + f_s(Q_h^{(t)}, \bar{\mu}_h^{(t)}, \nu_{h,\tau}^{\star}) - f_s(Q_{h,\tau}^{\star}, \mu_{h,\tau}^{\star}, \nu_{h,\tau}^{\star}) \\
&\le \Big(f_s(Q_h^{(t)}, \bar{\mu}_h^{(t)}, \bar{\nu}_h^{(t)}) - f_s(Q_h^{(t)}, \bar{\mu}_h^{(t)}, \nu_{h,\tau}^{\star})\Big) + f_s(Q_\tau^{\star}, \bar{\mu}^{(t)}, \nu_{h,\tau}^{\star}) - f_s(Q_{h,\tau}^{\star}, \mu_{h,\tau}^{\star}, \nu_{h,\tau}^{\star}) \\
&\qquad + \big\|Q_h^{(t)}(s) - Q_{h,\tau}^{\star}(s)\big\|_\infty \\
&\le f_s(Q_h^{(t)}, \bar{\mu}_h^{(t)}, \bar{\nu}_h^{(t)}) - f_s(Q_h^{(t)}, \bar{\mu}_h^{(t)}, \nu_{h,\tau}^{\star}) + \big\|Q_h^{(t)}(s) - Q_{h,\tau}^{\star}(s)\big\|_\infty.
\end{aligned}
$$

Note that Lemma 29 can be applied to the episodic setting by simply replacing $1/(1-\gamma)$ with H, which yields

$$
\begin{aligned}
f_s(Q_h^{(t)}, \bar{\mu}_h^{(t)}, \bar{\nu}_h^{(t)}) - f_s(Q_{h,\tau}^{\star}, \mu_{h,\tau}^{\star}, \nu_{h,\tau}^{\star}) \le{}& \big\|Q_h^{(t)}(s) - Q_{h,\tau}^{\star}(s)\big\|_\infty + 2\eta H\big\|Q_h^{(t)}(s) - Q_h^{(t-1)}(s)\big\|_\infty \\
&+ \frac{1-\eta\tau}{\eta}\mathsf{KL}_s\big(\nu_{h,\tau}^{\star} \,\|\, \nu_h^{(t-1)}\big) - \frac{1}{\eta}\mathsf{KL}_s\big(\nu_{h,\tau}^{\star} \,\|\, \nu_h^{(t)}\big) \\
&- \frac{1}{\eta}\big(1-4\eta H\big)\mathsf{KL}_s\big(\nu_h^{(t)} \,\|\, \bar{\nu}_h^{(t)}\big) - \frac{1-\eta\tau}{\eta}\mathsf{KL}_s\big(\bar{\nu}_h^{(t)} \,\|\, \nu_h^{(t-1)}\big) \\
&+ 2H\Big(\mathsf{KL}_s\big(\bar{\mu}_h^{(t)} \,\|\, \mu_h^{(t-1)}\big) + \mathsf{KL}_s\big(\mu_h^{(t-1)} \,\|\, \bar{\mu}_h^{(t-1)}\big)\Big).
\end{aligned} \tag{D.42}
$$

By a similar argument,

$$
\begin{aligned}
f_s(Q_{h,\tau}^{\star}, \mu_{h,\tau}^{\star}, \nu_{h,\tau}^{\star}) - f_s(Q_h^{(t)}, \bar{\mu}_h^{(t)}, \bar{\nu}_h^{(t)}) \le{}& \big\|Q_h^{(t)}(s) - Q_{h,\tau}^{\star}(s)\big\|_\infty + 2\eta H\big\|Q_h^{(t)}(s) - Q_h^{(t-1)}(s)\big\|_\infty \\
&+ \frac{1-\eta\tau}{\eta}\mathsf{KL}_s\big(\mu_{h,\tau}^{\star} \,\|\, \mu_h^{(t-1)}\big) - \frac{1}{\eta}\mathsf{KL}_s\big(\mu_{h,\tau}^{\star} \,\|\, \mu_h^{(t)}\big) \\
&- \frac{1}{\eta}\big(1-4\eta H\big)\mathsf{KL}_s\big(\mu_h^{(t)} \,\|\, \bar{\mu}_h^{(t)}\big) - \frac{1-\eta\tau}{\eta}\mathsf{KL}_s\big(\bar{\mu}_h^{(t)} \,\|\, \mu_h^{(t-1)}\big) \\
&+ 2H\Big(\mathsf{KL}_s\big(\bar{\nu}_h^{(t)} \,\|\, \nu_h^{(t-1)}\big) + \mathsf{KL}_s\big(\nu_h^{(t-1)} \,\|\, \bar{\nu}_h^{(t-1)}\big)\Big).
\end{aligned} \tag{D.43}
$$

Combining (D.42) + $\frac{2}{3}\cdot$ (D.43) gives

$$
\begin{aligned}
&\frac{1}{3}\big[f_s(Q_h^{(t)},\bar{\mu}_h^{(t)},\bar{\nu}_h^{(t)})-f_s(Q_{h,\tau}^\star,\mu_{h,\tau}^\star,\nu_{h,\tau}^\star)\big]\\
&\le\frac{5}{3}\big[\big\|Q_h^{(t)}(s)-Q_{h,\tau}^\star(s)\big\|_\infty+2\eta H\big\|Q_h^{(t)}(s)-Q_h^{(t-1)}(s)\big\|_\infty\big]\\
&\quad+\frac{1-\eta\tau}{\eta}\Big[\mathsf{KL}_s\big(\nu_{h,\tau}^\star\,\|\,\nu_h^{(t-1)}\big)+\frac{2}{3}\mathsf{KL}_s\big(\mu_{h,\tau}^\star\,\|\,\mu_h^{(t-1)}\big)\Big]-\frac{1}{\eta}\Big[\mathsf{KL}_s\big(\nu_{h,\tau}^\star\,\|\,\nu_h^{(t)}\big)+\frac{2}{3}\mathsf{KL}_s\big(\mu_{h,\tau}^\star\,\|\,\mu_h^{(t)}\big)\Big]\\
&\quad+2H\Big[\mathsf{KL}_s\big(\mu_h^{(t-1)}\,\|\,\bar{\mu}_h^{(t-1)}\big)+\frac{2}{3}\mathsf{KL}_s\big(\nu_h^{(t-1)}\,\|\,\bar{\nu}_h^{(t-1)}\big)\Big]\\
&\quad-\frac{1}{\eta}(1-4\eta H)\Big[\frac{2}{3}\mathsf{KL}_s\big(\mu_h^{(t)}\,\|\,\bar{\mu}_h^{(t)}\big)+\mathsf{KL}_s\big(\nu_h^{(t)}\,\|\,\bar{\nu}_h^{(t)}\big)\Big]\\
&\quad+\Big(2H-\frac{1-\eta\tau}{\eta}\cdot\frac{2}{3}\Big)\mathsf{KL}_s\big(\bar{\mu}^{(t)}\,\|\,\mu^{(t-1)}\big)+\Big(2H\cdot\frac{2}{3}-\frac{1-\eta\tau}{\eta}\Big)\mathsf{KL}_s\big(\bar{\nu}^{(t)}\,\|\,\nu^{(t-1)}\big).
\end{aligned}
\tag{D.44}
$$

With $\eta\le\frac{1}{8H}$, we have

$$
2H-\frac{1-\eta\tau}{\eta}\cdot\frac{2}{3}\le 0,\quad 2H\cdot\frac{2}{3}-\frac{1-\eta\tau}{\eta}\le 0,\quad\text{and}\quad\frac{1}{\eta}(1-\eta\tau)(1-4\eta H)\cdot\frac{2}{3}\ge 2H.
$$

Let

$$
G_h^{(t)}(s)=\mathsf{KL}_s\big(\nu_{h,\tau}^\star\,\|\,\nu_h^{(t)}\big)+\frac{2}{3}\mathsf{KL}_s\big(\mu_{h,\tau}^\star\,\|\,\mu_h^{(t)}\big)+\frac{2}{3}(1-4\eta H)\Big[\mathsf{KL}_s\big(\mu_h^{(t)}\,\|\,\bar{\mu}_h^{(t)}\big)+\mathsf{KL}_s\big(\nu_h^{(t)}\,\|\,\bar{\nu}_h^{(t)}\big)\Big].
$$

We can simplify (D.44) as

$$
\begin{aligned}
&f_s(Q_h^{(t)},\bar{\mu}_h^{(t)},\bar{\nu}_h^{(t)})-f_s(Q_{h,\tau}^\star,\mu_{h,\tau}^\star,\nu_{h,\tau}^\star)\\
&\le 5\big[\big\|Q_h^{(t)}(s)-Q_{h,\tau}^\star(s)\big\|_\infty+2\eta H\big\|Q_h^{(t)}(s)-Q_h^{(t-1)}(s)\big\|_\infty\big]+\frac{1-\eta\tau}{\eta}G_h^{(t-1)}(s)-\frac{1}{\eta}G_h^{(t)}(s).
\end{aligned}
$$

Plugging the above inequality into (D.41) gives

$$
\begin{aligned}
&Q_{h-1}^{(t_2)}(s,a,b)-Q_{h-1,\tau}^\star(s,a,b)\\
&\le(1-\eta\tau)^{t_2-t_1}2H\\
&\quad+\mathop{\mathbb{E}}_{s'\sim P_{h-1}(\cdot|s,a,b)}\left[5\eta\tau\sum_{l=t_1}^{t_2-1}(1-\eta\tau)^{t_2-1-l}\big(\big\|Q_h^{(l)}(s')-Q_{h,\tau}^\star(s')\big\|_\infty+2\eta H\big\|Q_h^{(l)}(s')-Q_h^{(l-1)}(s')\big\|_\infty\big)\right]\\
&\quad+\mathop{\mathbb{E}}_{s'\sim P_{h-1}(\cdot|s,a,b)}\Big[\tau(1-\eta\tau)^{t_2-t_1}G_h^{(t_1-1)}(s')\Big]\\
&\le(1-\eta\tau)^{t_2-t_1}2H\\
&\quad+10\eta\tau\mathop{\mathbb{E}}_{s'\sim P_{h-1}(\cdot|s,a,b)}\left[\sum_{l=t_1-1}^{t_2-1}(1-\eta\tau)^{t_2-1-l}\big\|Q_h^{(l)}(s')-Q_{h,\tau}^\star(s')\big\|_\infty\right]\\
&\quad+\tau(1-\eta\tau)^{t_2-t_1}\mathop{\mathbb{E}}_{s'\sim P_{h-1}(\cdot|s,a,b)}\Big[\mathsf{KL}_{s'}\big(\zeta_{h,\tau}^\star\,\|\,\zeta_h^{(t_1-1)}\big)+(1-4\eta H)\mathsf{KL}_{s'}\big(\zeta_h^{(t_1-1)}\,\|\,\bar{\zeta}_h^{(t_1-1)}\big)\Big].
\end{aligned}
$$

The other side of Lemma 26 can be shown with a similar proof and is therefore omitted.

D.5 Proof of auxiliary lemmas

D.5.1 Proof of Lemma 27

We first single out a set of bounds for $V^{(t)}$ and $Q^{(t)}$, which can be obtained by a simple induction:

$$\forall (s,a,b) \in \mathcal{S}\times\mathcal{A}\times\mathcal{B}, \qquad \begin{cases} -\frac{\tau\log|\mathcal{B}|}{1-\gamma} \le V^{(t)}(s) \le \frac{1+\tau\log|\mathcal{A}|}{1-\gamma} \\ -\frac{\gamma\tau\log|\mathcal{B}|}{1-\gamma} \le Q^{(t)}(s,a,b) \le \frac{1+\gamma\tau\log|\mathcal{A}|}{1-\gamma} \end{cases}. \tag{D.45}$$

We invoke the following lemma to bound several key quantities that will be helpful in the analysis.

Lemma 31 ([Mei et al., 2020b, Lemma 24]). *Let $\pi,\pi' \in \Delta(\mathcal{A})$ such that $\pi(a) \propto \exp(\theta(a))$, $\pi'(a)\propto\exp(\theta'(a))$ for some $\theta,\theta'\in\mathbb{R}^{|\mathcal{A}|}$. It holds that*

$$\big\|\pi-\pi'\big\|_1 \le \big\|\theta-\theta'\big\|_\infty.$$

With this lemma in mind, for any $t\ge 0$, it follows that

$$\begin{aligned}\big\|\bar\mu^{(t+1)}(s)-\mu^{(t+1)}(s)\big\|_1 &\le \min_{c\in\mathbb{R}}\big\|\log\bar\mu^{(t+1)}(s)-\log\mu^{(t+1)}(s)-c\cdot\mathbf{1}\big\|_\infty \\ &\le \eta\big\|Q^{(t)}(s)\bar\nu^{(t)}(s)-Q^{(t+1)}(s)\bar\nu^{(t+1)}(s)\big\|_\infty \\ &\le \eta\cdot\frac{1+\gamma\tau(\log|\mathcal{A}|+\log|\mathcal{B}|)}{1-\gamma} \le \frac{2\eta}{1-\gamma},\end{aligned}$$

where the second line follows from the update rule (5.11), and the last line follows from (D.45). A similar argument reveals that

$$\big\|\bar\nu^{(t+1)}(s)-\nu^{(t+1)}(s)\big\|_1 \le \frac{2\eta}{1-\gamma},$$

which completes the proof of (D.15a).

Moving onto the second claim (D.15b), we make note of the fact that when $t\ge 1$,

$$\begin{aligned}\bar\mu^{(t+1)}(a|s) &\propto \mu^{(t)}(a|s)^{1-\eta\tau}\exp(\eta[Q^{(t)}(s)\bar\nu^{(t)}(s)]_a) \\ &\overset{\text{(i)}}{\propto} \bar\mu^{(t)}(a|s)^{1-\eta\tau}\exp\Big(\eta\big[Q^{(t)}(s)\bar\nu^{(t)}(s)+(1-\eta\tau)(Q^{(t)}(s)\bar\nu^{(t)}(s)-Q^{(t-1)}(s)\bar\nu^{(t-1)}(s))\big]_a\Big) \\ &\propto \bar\mu^{(t)}(a|s)\exp(\eta w^{(t)}(a)),\end{aligned} \tag{D.46}$$

where

$$w^{(t)} = Q^{(t)}(s)\bar\nu^{(t)}(s)+(1-\eta\tau)\big(Q^{(t)}(s)\bar\nu^{(t)}(s)-Q^{(t-1)}(s)\bar\nu^{(t-1)}(s)\big)-\tau\log\bar\mu^{(t)}(s).$$

Here, (i) follows from the update rule (5.11) as

$$\begin{aligned}\mu^{(t)}(a|s) &\propto \mu^{(t-1)}(a|s)^{1-\eta\tau}\exp(\eta[Q^{(t)}(s)\bar\nu^{(t)}(s)]_a) \\ &\propto \mu^{(t-1)}(a|s)^{1-\eta\tau}\exp(\eta[Q^{(t-1)}(s)\bar\nu^{(t-1)}(s)]_a)\exp(\eta[Q^{(t)}(s)\bar\nu^{(t)}(s)-Q^{(t-1)}(s)\bar\nu^{(t-1)}(s)]_a) \\ &\propto \bar\mu^{(t)}(a|s)\exp(\eta[Q^{(t)}(s)\bar\nu^{(t)}(s)-Q^{(t-1)}(s)\bar\nu^{(t-1)}(s)]_a).\end{aligned}$$

Moreover, $w^{(t)}$ satisfies

$$\begin{aligned}\big\|w^{(t)}\big\|_\infty &\le \big\|Q^{(t)}(s)\bar\nu^{(t)}(s)\big\|_\infty+\big\|\tau\log\bar\mu^{(t)}(s)\big\|_\infty+(1-\eta\tau)\big\|Q^{(t)}(s)\bar\nu^{(t)}(s)-Q^{(t-1)}(s)\bar\nu^{(t-1)}(s)\big\|_\infty \\ &\le \frac{2}{1-\gamma}+\frac{2}{1-\gamma}+\frac{2(1-\eta\tau)}{1-\gamma} \le \frac{6}{1-\gamma}.\end{aligned}$$

Here, the second step is due to (D.16), which we shall prove momentarily. Recall that when $t = 0$, we have $\bar{\mu}^{(t+1)} = \bar{\mu}^{(0)}$. In sum, we have

$$\forall s \in \mathcal{S}, t \geq 0, \qquad \left\|\bar{\mu}^{(t+1)}(s) - \bar{\mu}^{(t)}(s)\right\|_1 \leq \frac{6\eta}{1-\gamma},$$

concluding the proof of (D.15b).

It remains to prove the claim (D.16). For simplicity we focus on the bound with $\left\|\log \mu^{(t)}(s)\right\|_\infty$; the other bounds follow similarly. It is worth noting that $\mu^{(t)}(s)$ can be always written as $\mu^{(t)}(a|s) \propto \exp(w^{(t)}(a)/\tau)$ for some $w^{(t)} \in \mathbb{R}^{|\mathcal{A}|}$ satisfying

$$\forall a \in \mathcal{A}, \qquad -\frac{\gamma\tau \log|\mathcal{B}|}{1-\gamma} \leq w^{(t)}(a) \leq \frac{1+\gamma\tau\log|\mathcal{A}|}{1-\gamma}.$$

To see this, note that the claim trivially holds for $t = 0$ with $w^{(0)} = \mathbf{0}$. When the statement holds for some $t \geq 0$, we have

$$\begin{aligned}
\mu^{(t+1)}(a|s) &\propto \mu^{(t)}(a|s)^{1-\eta\tau} \exp(\eta Q^{(t+1)}(s)\bar{\nu}^{(t+1)}(s)) \\
&\propto \exp\left(((1-\eta\tau)w^{(t)} + \eta\tau Q^{(t+1)}(s)\bar{\nu}^{(t+1)}(s))/\tau\right) \\
&\propto \exp\left(w^{(t+1)}/\tau\right),
\end{aligned}$$

with $w^{(t+1)} = (1-\eta\tau)w^{(t)} + \eta\tau Q^{(t+1)}(s)\bar{\nu}^{(t+1)}(s)$. We conclude that the claim holds for $t+1$ by recalling (D.45). It then follows straightforwardly that

$$\frac{\mu^{(t)}(a_1|s)}{\mu^{(t)}(a_2|s)} = \exp\left(\frac{w^{(t)}(a_1) - w^{(t)}(a_2)}{\tau}\right) \leq \exp\left(\frac{1+\gamma\tau(\log|\mathcal{A}| + \log|\mathcal{B}|)}{(1-\gamma)\tau}\right)$$

for any $a_1, a_2 \in \mathcal{A}$. This allows us to show that

$$\min_{a\in\mathcal{A}} \mu^{(t)}(a|s) \geq \frac{1}{|\mathcal{A}|\exp\left(\frac{1+\gamma\tau(\log|\mathcal{A}|+\log|\mathcal{B}|)}{(1-\gamma)\tau}\right)} \sum_{a\in\mathcal{A}} \mu^{(t)}(a|s) = \frac{1}{|\mathcal{A}|\exp\left(\frac{1+\gamma\tau(\log|\mathcal{A}|+\log|\mathcal{B}|)}{(1-\gamma)\tau}\right)},$$

which gives

$$\begin{aligned}
\|\log \mu^{(t)}(s)\|_\infty &\leq \frac{1+\gamma\tau(\log|\mathcal{A}| + \log|\mathcal{B}|)}{(1-\gamma)\tau} + \log|\mathcal{A}| \leq \frac{1}{(1-\gamma)\tau} + \frac{\log|\mathcal{A}| + \gamma\log|\mathcal{B}|}{1-\gamma} \\
&\leq \frac{2}{(1-\gamma)\tau}.
\end{aligned}$$

D.5.2 Proof of Lemma 28

We decompose the term $f_s(Q^{(t+1)}, \bar{\mu}^{(t+1)}, \bar{\nu}^{(t+1)}) - f_s(Q^{(t)}, \bar{\mu}^{(t)}, \bar{\nu}^{(t)})$ as follows:

$$\begin{aligned}
&f_s(Q^{(t+1)}, \bar{\mu}^{(t+1)}, \bar{\nu}^{(t+1)}) - f_s(Q^{(t)}, \bar{\mu}^{(t)}, \bar{\nu}^{(t)}) \\
&= f_s(Q^{(t+1)}, \bar{\mu}^{(t+1)}, \bar{\nu}^{(t+1)}) - f_s(Q^{(t)}, \bar{\mu}^{(t+1)}, \bar{\nu}^{(t+1)}) + f_s(Q^{(t)}, \bar{\mu}^{(t+1)}, \bar{\nu}^{(t+1)}) - f_s(Q^{(t)}, \bar{\mu}^{(t)}, \bar{\nu}^{(t)}) \\
&= \bar{\mu}^{(t+1)}(s)^\top \Big(Q^{(t+1)}(s) - Q^{(t)}(s)\Big)\bar{\nu}^{(t+1)}(s) \\
&\quad + f_s(Q^{(t)}, \bar{\mu}^{(t+1)}, \bar{\nu}^{(t)}) - f_s(Q^{(t)}, \bar{\mu}^{(t)}, \bar{\nu}^{(t)}) + f_s(Q^{(t)}, \bar{\mu}^{(t)}, \bar{\nu}^{(t+1)}) - f_s(Q^{(t)}, \bar{\mu}^{(t)}, \bar{\nu}^{(t)}) \\
&\quad + \Big[f_s(Q^{(t)}, \bar{\mu}^{(t+1)}, \bar{\nu}^{(t+1)}) + f_s(Q^{(t)}, \bar{\mu}^{(t)}, \bar{\nu}^{(t)}) - f_s(Q^{(t)}, \bar{\mu}^{(t+1)}, \bar{\nu}^{(t)}) - f_s(Q^{(t)}, \bar{\mu}^{(t)}, \bar{\nu}^{(t+1)})\Big].
\end{aligned}$$

Note that $\big|\bar{\mu}^{(t+1)}(s)^\top\big(Q^{(t+1)}(s)-Q^{(t)}(s)\big)\bar{\nu}^{(t+1)}(s)\big| \le \big\|Q^{(t+1)}(s)-Q^{(t)}(s)\big\|_\infty$. For the terms in the bracket, we have

$$\begin{aligned}
&\Big|\Big[f_s(Q^{(t)},\bar{\mu}^{(t+1)},\bar{\nu}^{(t+1)})+f_s(Q^{(t)},\bar{\mu}^{(t)},\bar{\nu}^{(t)})-f_s(Q^{(t)},\bar{\mu}^{(t+1)},\bar{\nu}^{(t)})-f_s(Q^{(t)},\bar{\mu}^{(t)},\bar{\nu}^{(t+1)})\Big]\Big| \\
&= \Big|\big(\bar{\mu}^{(t+1)}(s)-\bar{\mu}^{(t)}(s)\big)^\top Q^{(t)}(s)\big(\bar{\nu}^{(t+1)}(s)-\bar{\nu}^{(t)}(s)\big)\Big| \\
&\le \frac{2}{1-\gamma}\mathsf{KL}_s\big(\bar{\zeta}^{(t+1)}\,\|\,\bar{\zeta}^{(t)}\big),
\end{aligned}$$

where the last step invokes Cauchy-Schwarz inequality and Pinsker's inequality (see e.g., (D.22)). It remains to bound the two difference terms $\big|f_s(Q^{(t)},\bar{\mu}^{(t+1)},\bar{\nu}^{(t)})-f_s(Q^{(t)},\bar{\mu}^{(t)},\bar{\nu}^{(t)})\big|$ and $\big|f_s(Q^{(t)},\bar{\mu}^{(t)},\bar{\nu}^{(t+1)})-f_s(Q^{(t)},\bar{\mu}^{(t)},\bar{\nu}^{(t)})\big|$. To proceed, we show that

$$\begin{aligned}
&f_s(Q^{(t)},\bar{\mu}^{(t)},\bar{\nu}^{(t)})-f_s(Q^{(t)},\bar{\mu}^{(t+1)},\bar{\nu}^{(t)}) \\
&= \Big\langle \bar{\mu}^{(t)}(s)-\bar{\nu}^{(t+1)}(s), Q^{(t)}(s)^\top\bar{\mu}^{(t)}(s)\Big\rangle + \tau\mathcal{H}(\bar{\mu}^{(t)}(s))-\tau\mathcal{H}(\bar{\mu}^{(t+1)}(s)) \\
&= \Big\langle \bar{\mu}^{(t)}(s)-\bar{\mu}^{(t+1)}(s), Q^{(t)}(s)^\top\bar{\nu}^{(t)}(s)+(1-\eta\tau)\big(Q^{(t)}(s)\bar{\nu}^{(t)}(s)-Q^{(t-1)}(s)\bar{\nu}^{(t-1)}(s)\big)\Big\rangle \\
&\quad +\tau\mathcal{H}(\bar{\mu}^{(t)}(s))-\tau\mathcal{H}(\bar{\mu}^{(t+1)}(s)) \\
&\quad -(1-\eta\tau)\Big\langle \bar{\mu}^{(t)}(s)-\bar{\mu}^{(t+1)}(s), Q^{(t)}(s)\bar{\nu}^{(t)}(s)-Q^{(t-1)}(s)\bar{\nu}^{(t-1)}(s)\Big\rangle \\
&= -\frac{1}{\eta}\mathsf{KL}_s\big(\bar{\mu}^{(t)}\,\|\,\bar{\mu}^{(t+1)}\big)-\frac{1-\eta\tau}{\eta}\mathsf{KL}_s\big(\bar{\mu}^{(t+1)}\,\|\,\bar{\mu}^{(t)}\big) \\
&\quad -(1-\eta\tau)\Big\langle \bar{\mu}^{(t)}(s)-\bar{\mu}^{(t+1)}(s), Q^{(t)}(s)\bar{\nu}^{(t)}(s)-Q^{(t-1)}(s)\bar{\nu}^{(t-1)}(s)\Big\rangle. \qquad \text{(D.47)}
\end{aligned}$$

Here, the third step results from the special case of the following three-point lemma—which is proven in Appendix D.5.5—in view of (D.46).

Lemma 32 (Regularized three-point lemma)**.** *Let $x\in\Delta(\mathcal{A})$ be defined as*

$$x(a)\propto y(a)^{1-\eta\tau}\exp(-\eta w(a))$$

for some $w\in\mathbb{R}^{|\mathcal{A}|}$ and $y\in\Delta(\mathcal{A})$. It holds for all $z\in\Delta(\mathcal{A})$ that

$$\frac{\eta}{1-\eta\tau}\Big[\langle x-z,w\rangle-\tau\mathcal{H}(x)+\tau\mathcal{H}(z)\Big]=\mathsf{KL}\left(z\,\|\,y\right)-\frac{1}{1-\eta\tau}\mathsf{KL}\left(z\,\|\,x\right)-\mathsf{KL}\left(x\,\|\,y\right).$$

This immediately implies that

$$\frac{\eta}{1-\eta\tau}\Big[\langle x-y,w\rangle-\tau\mathcal{H}(x)+\tau\mathcal{H}(y)\Big]=-\frac{1}{1-\eta\tau}\mathsf{KL}\left(y\,\|\,x\right)-\mathsf{KL}\left(x\,\|\,y\right).$$

Recall from the earlier discussion (cf. (D.46)) that $\bar{\mu}^{(t+1)}(a|s)\propto\bar{\mu}^{(t)}(a|s)\exp(\eta w^{(t)}(s))$ for some $w^{(t)}\in\mathbb{R}^{|\mathcal{B}|}$ satisfying

$$\big\|w^{(t)}\big\|_\infty\le\frac{6}{1-\gamma}.$$

We can ensure that $\|\eta w^{(t)}\|_\infty\le 1/30$ as long as $\eta^{-1}\ge\frac{180}{1-\gamma}$, and the next lemma guarantees $\mathsf{KL}_s\big(\bar{\mu}^{(t)}\,\|\,\bar{\mu}^{(t+1)}\big)\le 2\mathsf{KL}_s\big(\bar{\mu}^{(t+1)}\,\|\,\bar{\mu}^{(t)}\big)$ in this case.

Lemma 33. *Let $w\in\mathbb{R}^{|\mathcal{A}|}$, $\pi,\pi'\in\Delta(\mathcal{A})$ satisfy, for each $a\in\mathcal{A}$, $\pi'(a)\propto\pi(a)\exp(w(a))$ with $\|w\|_\infty\le\frac{1}{30}$. It holds that*

$$\mathsf{KL}\left(\pi\,\|\,\pi'\right)\le 2\mathsf{KL}\left(\pi'\,\|\,\pi\right).$$

Therefore, we can continue to bound (D.47) by

$$
\begin{aligned}
&\big|f_s(Q^{(t)},\bar{\mu}^{(t+1)},\bar{\nu}^{(t)})-f_s(Q^{(t)},\bar{\mu}^{(t)},\bar{\nu}^{(t)})\big| \\
&\le \frac{1}{\eta}\mathsf{KL}_s\big(\bar{\mu}^{(t)}\,\|\,\bar{\mu}^{(t+1)}\big)+\frac{1-\eta\tau}{\eta}\mathsf{KL}_s\big(\bar{\mu}^{(t+1)}\,\|\,\bar{\mu}^{(t)}\big) \\
&\quad+\big\|\bar{\mu}^{(t+1)}(s)-\bar{\mu}^{(t)}(s)\big\|_1\big\|Q^{(t)}(s)\bar{\nu}^{(t)}(s)-Q^{(t-1)}(s)\bar{\nu}^{(t-1)}(s)\big\|_\infty \\
&\le \frac{3}{\eta}\mathsf{KL}_s\big(\bar{\mu}^{(t+1)}\,\|\,\bar{\mu}^{(t)}\big)+\big\|\bar{\mu}^{(t+1)}(s)-\bar{\mu}^{(t)}(s)\big\|_1\big\|Q^{(t)}(s)-Q^{(t-1)}(s)\big\|_\infty \\
&\quad+\big\|Q^{(t)}(s)\big\|_\infty\big\|\bar{\mu}^{(t+1)}(s)-\bar{\mu}^{(t)}(s)\big\|_1\big\|\bar{\nu}^{(t)}(s)-\bar{\nu}^{(t-1)}(s)\big\|_1 \\
&\le \Big(\frac{3}{\eta}+\frac{2}{1-\gamma}\Big)\mathsf{KL}_s\big(\bar{\mu}^{(t+1)}\,\|\,\bar{\mu}^{(t)}\big)+\frac{2}{1-\gamma}\mathsf{KL}_s\big(\bar{\mu}^{(t)}\,\|\,\bar{\mu}^{(t-1)}\big)+\frac{6\eta}{1-\gamma}\big\|Q^{(t)}(s)-Q^{(t-1)}(s)\big\|_\infty,
\end{aligned}
$$

where the last line uses Lemma 27, Cauchy-Schwarz inequality and Pinsker's inequality (see e.g., (D.22)). One can bound $\big|f_s(Q^{(t)},\bar{\mu}^{(t)},\bar{\nu}^{(t)})-f_s(Q^{(t)},\bar{\mu}^{(t)},\bar{\nu}^{(t+1)})\big|$ with similar arguments. Putting all pieces together, we arrive at

$$
\begin{aligned}
&\big|f_s(Q^{(t+1)},\bar{\mu}^{(t+1)},\bar{\nu}^{(t+1)})-f_s(Q^{(t)},\bar{\mu}^{(t)},\bar{\nu}^{(t)})\big| \\
&\le \Big\|Q^{(t+1)}(s)-Q^{(t)}(s)\Big\|_\infty+\Big(\frac{3}{\eta}+\frac{4}{1-\gamma}\Big)\mathsf{KL}_s\big(\bar{\zeta}^{(t+1)}\,\|\,\bar{\zeta}^{(t)}\big)+\frac{2}{1-\gamma}\mathsf{KL}_s\big(\bar{\zeta}^{(t)}\,\|\,\bar{\zeta}^{(t-1)}\big) \\
&\quad+\frac{12\eta}{1-\gamma}\big\|Q^{(t)}(s)-Q^{(t-1)}(s)\big\|_\infty.
\end{aligned}
$$

D.5.3 Proof of Lemma 29

Note that

$$
\begin{aligned}
&f_s(Q^{(t)},\bar{\mu}^{(t)},\bar{\nu}^{(t)})-f_s(Q^{(t)},\bar{\mu}^{(t)},\nu) \\
&=\Big\langle\bar{\nu}^{(t)}(s)-\nu_\tau^\star(s),Q^{(t)}(s)^\top\bar{\mu}^{(t)}(s)\Big\rangle-\tau\mathcal{H}(\bar{\nu}^{(t)}(s))+\tau\mathcal{H}(\nu_\tau^\star(s)) \\
&=\Big\langle\bar{\nu}^{(t)}(s)-\nu^{(t)}(s),Q^{(t)}(s)^\top\bar{\mu}^{(t)}(s)-Q^{(t-1)}(s)^\top\bar{\mu}^{(t-1)}(s)\Big\rangle \\
&\quad+\Big\langle\bar{\nu}^{(t)}(s)-\nu^{(t)}(s),Q^{(t-1)}(s)^\top\bar{\mu}^{(t-1)}(s)\Big\rangle-\tau\mathcal{H}(\bar{\nu}^{(t)}(s))+\tau\mathcal{H}(\nu^{(t)}(s)) \\
&\quad+\Big\langle\nu^{(t)}(s)-\nu_\tau^\star(s),Q^{(t)}(s)^\top\bar{\mu}^{(t)}(s)\Big\rangle-\tau\mathcal{H}(\nu^{(t)}(s))+\tau\mathcal{H}(\nu_\tau^\star(s)) \\
&=\Big\langle\bar{\nu}^{(t)}(s)-\nu^{(t)}(s),Q^{(t)}(s)^\top\bar{\mu}^{(t)}(s)-Q^{(t-1)}(s)^\top\bar{\mu}^{(t-1)}(s)\Big\rangle \\
&\quad+\frac{1-\eta\tau}{\eta}\mathsf{KL}_s\big(\nu^{(t)}\,\|\,\nu^{(t-1)}\big)-\frac{1}{\eta}\mathsf{KL}_s\big(\nu^{(t)}\,\|\,\bar{\nu}^{(t)}\big)-\frac{1-\eta\tau}{\eta}\mathsf{KL}_s\big(\bar{\nu}^{(t)}\,\|\,\nu^{(t-1)}\big) \\
&\quad+\frac{1-\eta\tau}{\eta}\mathsf{KL}_s\big(\nu_\tau^\star\,\|\,\nu^{(t-1)}\big)-\frac{1}{\eta}\mathsf{KL}_s\big(\nu_\tau^\star\,\|\,\nu^{(t)}\big)-\frac{1-\eta\tau}{\eta}\mathsf{KL}_s\big(\nu^{(t)}\,\|\,\nu^{(t-1)}\big) \\
&\le\big\|\bar{\nu}^{(t)}(s)-\nu^{(t)}(s)\big\|_1\big\|Q^{(t)}(s)^\top\bar{\mu}^{(t)}(s)-Q^{(t-1)}(s)^\top\bar{\mu}^{(t-1)}(s)\big\|_\infty \\
&\quad-\frac{1}{\eta}\mathsf{KL}_s\big(\nu^{(t)}\,\|\,\bar{\nu}^{(t)}\big)-\frac{1-\eta\tau}{\eta}\mathsf{KL}_s\big(\bar{\nu}^{(t)}\,\|\,\nu^{(t-1)}\big)+\frac{1-\eta\tau}{\eta}\mathsf{KL}_s\big(\nu_\tau^\star\,\|\,\nu^{(t-1)}\big)-\frac{1}{\eta}\mathsf{KL}_s\big(\nu_\tau^\star\,\|\,\nu^{(t)}\big).
\end{aligned}
\tag{D.48}
$$

Here, the second step results from Lemma 32. We further bound the first term in (D.48) as follows.

$$
\begin{aligned}
&\big\|\bar{\nu}^{(t)}(s)-\nu^{(t)}(s)\big\|_1\big\|Q^{(t)}(s)^\top\bar{\mu}^{(t)}(s)-Q^{(t-1)}(s)^\top\bar{\mu}^{(t-1)}(s)\big\|_\infty\\
&\le\big\|\bar{\nu}^{(t)}(s)-\nu^{(t)}(s)\big\|_1\Big(\big\|\big(Q^{(t)}(s)-Q^{(t-1)}(s)\big)^\top\bar{\mu}^{(t-1)}(s)\big\|_\infty+\big\|Q^{(t)}(s)\big(\bar{\mu}^{(t)}(s)-\bar{\mu}^{(t-1)}(s)\big)\big\|_\infty\Big)\\
&\le\big\|\bar{\nu}^{(t)}(s)-\nu^{(t)}(s)\big\|_1\big\|Q^{(t)}(s)-Q^{(t-1)}(s)\big\|_\infty+\frac{2}{1-\gamma}\big\|\bar{\nu}^{(t)}(s)-\nu^{(t)}(s)\big\|_1\big\|\bar{\mu}^{(t)}(s)-\bar{\mu}^{(t-1)}(s)\big\|_1\\
&\le\frac{2\eta}{1-\gamma}\big\|Q^{(t)}(s)-Q^{(t-1)}(s)\big\|_\infty+\frac{1}{1-\gamma}\Big[2\big\|\bar{\nu}^{(t)}(s)-\nu^{(t)}(s)\big\|_1^2\\
&\qquad+\big\|\bar{\mu}^{(t)}(s)-\mu^{(t-1)}(s)\big\|_1^2+\big\|\mu^{(t-1)}(s)-\bar{\mu}^{(t-1)}(s)\big\|_1^2\Big]\\
&\le\frac{2\eta}{1-\gamma}\big\|Q^{(t)}(s)-Q^{(t-1)}(s)\big\|_\infty+\frac{4}{1-\gamma}\mathsf{KL}_s\big(\nu^{(t)}\,\|\,\bar{\nu}^{(t)}\big)\\
&\qquad+\frac{2}{1-\gamma}\mathsf{KL}\left(\bar{\mu}^{(t)}(s)\,\|\,\mu^{(t-1)}(s)\right)+\frac{2}{1-\gamma}\mathsf{KL}\left(\mu^{(t-1)}(s)\,\|\,\bar{\mu}^{(t-1)}(s)\right),
\end{aligned}
$$

where the penultimate inequality follows from Lemma 27, and the last line follows from Pinsker's inequality. Substitution of the above inequality into (D.48) completes the proof.

D.5.4 Proof of Lemma 30

By definition, we have

$$
\begin{aligned}
\delta_{l,t}&=\alpha_l\prod_{i=l+1}^{t}(1-c_1\alpha_i)\\
&=\alpha_l\prod_{i=l+1}^{t}(1-c_2\alpha_i+(c_2-c_1)\alpha_i)\\
&=\alpha_l(c_2-c_1)\alpha_{l+1}\prod_{i=l+2}^{t}(1-c_2\alpha_i+(c_2-c_1)\alpha_i)+\alpha_l(1-c_2\alpha_{l+1})\prod_{i=l+2}^{t}(1-c_2\alpha_i+(c_2-c_1)\alpha_i).
\end{aligned}
$$

Continuing this expansion recursively, we obtain

$$
\begin{aligned}
\delta_{l,t}&=\alpha_l\sum_{i=l+1}^{t}(c_2-c_1)\alpha_i\cdot\prod_{j=l+1}^{i}(1-c_2\alpha_j)\cdot\prod_{k=i+1}^{t}(1-c_1\alpha_k)+\alpha_l\prod_{i=l+1}^{t}(1-c_2\alpha_i)\\
&=(c_2-c_1)\sum_{i=l+1}^{t}\xi_{l,i}\delta_{i,t}+\xi_{l,t}.
\end{aligned}
$$

Rearranging terms, it follows that

$$
\begin{aligned}
\sum_{i=l}^{t}\xi_{l,i}\delta_{i+1,t}&=\alpha_l\delta_{l+1,t}+\sum_{i=l+1}^{t}\xi_{l,i}\delta_{i+1,t}\\
&=\frac{\alpha_{l+1}}{1-c_1\alpha_{l+1}}\delta_{l,t}+\sum_{i=l+1}^{t}\xi_{l,i}\delta_{i,t}\cdot\frac{\alpha_{i+1}}{\alpha_i(1-c_1\alpha_{i+1})}\\
&\overset{\text{(i)}}{\le}\delta_{l,t}+2\sum_{i=l+1}^{t}\xi_{l,i}\delta_{i,t}=\delta_{l,t}+\frac{2}{c_2-c_1}(\delta_{l,t}-\xi_{l,t})\le\Big(1+\frac{2}{c_2-c_1}\Big)\delta_{l,t},
\end{aligned}
$$

where the second line results from the definition of $\delta_{l,t}$ and (i) is due to $\{\alpha_i\}$ being non-increasing and

$$\alpha_{l+1} \le \eta\tau \le 1/2, \quad 1 - c_1\alpha_l \ge 1/2$$

for all $l \ge 1$.

D.5.5 Proof of Lemma 32

We have

$$\begin{aligned}
\mathsf{KL}\left(z \,\|\, y\right) &= -\mathcal{H}(z) + \mathcal{H}(y) - \langle z - y, \log y\rangle \\
&= -\mathcal{H}(z) + \mathcal{H}(x) - \langle z - x, \log y\rangle - \mathcal{H}(x) + \mathcal{H}(y) - \langle x - y, \log y\rangle \\
&= -\mathcal{H}(z) + \mathcal{H}(x) - \langle z - x, \log x\rangle - \mathcal{H}(x) + \mathcal{H}(y) - \langle x - y, \log y\rangle - \langle z - x, \log y - \log x\rangle \\
&= \mathsf{KL}\left(z \,\|\, x\right) + \mathsf{KL}\left(x \,\|\, y\right) - \frac{\eta}{1-\eta\tau}\left\langle z - x, w + \tau \log x\right\rangle,
\end{aligned}$$

where the last line follows from the update rule. Rearranging terms gives

$$\frac{\eta}{1-\eta\tau}\left\langle x - z, w\right\rangle = \mathsf{KL}\left(z \,\|\, y\right) - \mathsf{KL}\left(z \,\|\, x\right) - \mathsf{KL}\left(x \,\|\, y\right) + \frac{\eta\tau}{1-\eta\tau}\left\langle z - x, \log x\right\rangle.$$

Adding $\frac{\eta\tau}{1-\eta\tau}(-\mathcal{H}(x) + \mathcal{H}(z))$ to both sides, we are left with

$$\begin{aligned}
\frac{\eta}{1-\eta\tau}\Big[\left\langle x - z, w\right\rangle - \tau\mathcal{H}(x) + \tau\mathcal{H}(z)\Big] &= \mathsf{KL}\left(z \,\|\, y\right) - \mathsf{KL}\left(z \,\|\, x\right) - \mathsf{KL}\left(x \,\|\, y\right) \\
&\quad - \frac{\eta\tau}{1-\eta\tau}\big(-\mathcal{H}(z) + \mathcal{H}(x) - \langle z - x, \log x\rangle\big) \\
&= \mathsf{KL}\left(z \,\|\, y\right) - \frac{1}{1-\eta\tau}\mathsf{KL}\left(z \,\|\, x\right) - \mathsf{KL}\left(x \,\|\, y\right).
\end{aligned}$$

D.5.6 Proof of Lemma 33

We begin with a simple sandwich bound of $\log(1+x)$ which will be used later: when $x > -\frac{1}{10}$, we have

$$x - \Big(\frac{1}{2} + \frac{|x|}{2}\Big)x^2 \le \log(1+x) \le x - \Big(\frac{1}{2} - \frac{|x|}{3}\Big)x^2. \tag{D.49}$$

We shall prove this at the end of this proof. The following lemma, which is standard (see, e.g., [Mei et al., 2020b, Lemma 23], [Cen et al., 2022b, Lemma 3]), allows us to control $\big\|\log\pi - \log\pi'\big\|_\infty$, and in turn $\|\pi/\pi'\|_\infty$.

Lemma 34. *Let $\pi, \pi' \in \Delta(\mathcal{A})$ satisfy $\pi(a) \propto \exp(\theta(a))$ and $\pi'(a) \propto \exp(\theta'(a))$ for some $\theta, \theta' \in \mathbb{R}^{|\mathcal{A}|}$. It holds that*

$$\big\|\log\pi - \log\pi'\big\|_\infty \le 2\big\|\theta - \theta'\big\|_\infty.$$

In view of the above lemma, and since $\big\|w\big\|_\infty < 1/30$, we have $\forall a \in \mathcal{A}$:

$$\begin{aligned}
\Big|\frac{\pi(a)}{\pi'(a)} - 1\Big| &= \Big|\exp\Big(\log\frac{\pi(a)}{\pi'(a)}\Big) - \exp(0)\Big| \le |\log\pi(a) - \log\pi'(a)|\max\Big\{1, \frac{\pi(a)}{\pi'(a)}\Big\} \\
&\le 2\big\|w\big\|_\infty \exp(2\big\|w\big\|_\infty) \le 3\big\|w\big\|_\infty.
\end{aligned} \tag{D.50}$$

Therefore, we can bound $\mathsf{KL}\left(\pi \,\|\, \pi'\right)$ as

$$\begin{aligned}
\mathsf{KL}\left(\pi \,\|\, \pi'\right) &= \sum_{a\in\mathcal{A}} \pi(a) \log \frac{\pi(a)}{\pi'(a)} \\
&\overset{\text{(i)}}{\leq} \sum_{a\in\mathcal{A}} \pi(a) \left(\frac{\pi(a)}{\pi'(a)} - 1 - \Big(\frac{1}{2} - \big\|w\big\|_\infty\Big)\Big(\frac{\pi(a)}{\pi'(a)} - 1\Big)^2 \right) \\
&\overset{\text{(ii)}}{=} \sum_{a\in\mathcal{A}} \big(\pi(a) - \pi'(a)\big)\Big(\frac{\pi(a)}{\pi'(a)} - 1\Big) + \sum_{a\in\mathcal{A}} \pi'(a)\Big(\frac{\pi(a)}{\pi'(a)} - 1\Big) - \Big(\frac{1}{2} - \big\|w\big\|_\infty\Big) \sum_{a\in\mathcal{A}} \pi(a)\Big(\frac{\pi(a)}{\pi'(a)} - 1\Big)^2 \\
&= \chi^2(\pi;\pi') - \Big(\frac{1}{2} - \big\|w\big\|_\infty\Big) \sum_{a\in\mathcal{A}} \pi(a)\Big(\frac{\pi(a)}{\pi'(a)} - 1\Big)^2 \\
&\overset{\text{(iii)}}{\leq} \chi^2(\pi;\pi') - \Big(\frac{1}{2} - \big\|w\big\|_\infty\Big)\big(1 - 3\big\|w\big\|_\infty\big) \sum_{a\in\mathcal{A}} \pi'(a)\Big(\frac{\pi(a)}{\pi'(a)} - 1\Big)^2 \\
&= \left(1 - \Big(\frac{1}{2} - \big\|w\big\|_\infty\Big)\big(1 - 3\big\|w\big\|_\infty\big)\right) \chi^2(\pi;\pi'),
\end{aligned} \tag{D.51}$$

where (i) follows from (D.49), (ii) utilizes the fact $\sum_{a\in\mathcal{A}}(\pi(a) - \pi'(a)) = 0$, and (iii) makes use of (D.50). On the other hand, by similar arguments, we have

$$\begin{aligned}
\mathsf{KL}\left(\pi' \,\|\, \pi\right) &= \sum_{a\in\mathcal{A}} \pi'(a) \log \frac{\pi'(a)}{\pi(a)} \\
&\geq \sum_{a\in\mathcal{A}} \pi'(a) \left(\frac{\pi'(a)}{\pi(a)} - 1 - \frac{(1 + 3\big\|w\big\|_\infty)}{2}\Big(\frac{\pi'(a)}{\pi(a)} - 1\Big)^2 \right) \\
&= \chi^2(\pi';\pi) - \frac{(1 + 3\big\|w\big\|_\infty)}{2} \sum_{a\in\mathcal{A}} \pi'(a)\Big(\frac{\pi'(a)}{\pi(a)} - 1\Big)^2 \\
&\geq \chi^2(\pi';\pi) - \frac{(1 + 3\big\|w\big\|_\infty)^2}{2} \sum_{a\in\mathcal{A}} \pi(a)\Big(\frac{\pi'(a)}{\pi(a)} - 1\Big)^2 \\
&= \Big(1 - \frac{(1 + 3\big\|w\big\|_\infty)^2}{2}\Big)\chi^2(\pi';\pi).
\end{aligned} \tag{D.52}$$

By definition of $\chi^2(\pi;\pi')$, we further have

$$\begin{aligned}
\chi^2(\pi;\pi') &= \sum_{a\in\mathcal{A}} \pi'(a)\Big(\frac{\pi(a)}{\pi'(a)} - 1\Big)^2 \\
&\leq \big\|\pi/\pi'\big\|_\infty \sum_{a\in\mathcal{A}} \frac{\big(\pi'(a) - \pi(a)\big)^2}{\pi(a)} \\
&\leq (1 + 3\big\|w\big\|_\infty)\chi^2(\pi';\pi),
\end{aligned} \tag{D.53}$$

where the last line uses (D.50). Combining (D.51), (D.52) and (D.53) gives

$$\mathsf{KL}\left(\pi \,\|\, \pi'\right) \leq (1 + 3\big\|w\big\|_\infty) \cdot \left[\frac{1 - \big(1/2 - \big\|w\big\|_\infty\big)(1 - 3\big\|w\big\|_\infty)}{1 - (1 + 3\big\|w\big\|_\infty)^2/2}\right] \mathsf{KL}\left(\pi' \,\|\, \pi\right).$$

It is straightforward to verify that the factor is less than 2 when $\big\|w\big\|_\infty \leq 1/30$.

Proof of (D.49)**.** For any $x > -1$, it holds that

$$\begin{aligned}\log(1+x) &\le x - \frac{x^2}{2} + \frac{x^3}{3} \\ &\le x - \frac{x^2}{2} + \frac{|x^3|}{3} = x - \Big(\frac{1}{2} - \frac{|x|}{3}\Big)x^2,\end{aligned}$$

and that

$$\begin{aligned}\log(1+x) &\ge x - \frac{x^2}{2} + \frac{x^3}{3(1+x)^3} \\ &\ge x - \frac{x^2}{2} - \frac{|x^3|}{3(1+x)^3} = x - \Big(\frac{1}{2} + \frac{|x|}{3(1+x)^3}\Big)x^2.\end{aligned}$$

Therefore, when $x > -\frac{1}{10}$, we have $(1+x)^3 > \frac{2}{3}$ and thus

$$x - \Big(\frac{1}{2} + \frac{|x|}{2}\Big)x^2 \le \log(1+x) \le x - \Big(\frac{1}{2} - \frac{|x|}{3}\Big)x^2.$$

Appendix E

Proofs for Chapter 6

E.1 Proof for single-timescale OMWU (Section 6.2)

Before delving into the main proof, we first record a useful lemma pertaining to a basic property of zero-sum polymatrix games; the proof is deferred to Appendix E.3.1. For $i \in V$, we denote by $\mathcal{N}_i = \{j : (i,j) \in E\}$ the neighbors of agent i in the graph (V, E). For notational simplicity, we denote by $x \overset{\mathbf{1}}{=} y$ the equivalence between two vectors x and y up to a global shift, i.e.,

$$x = y + c \cdot \mathbf{1} \tag{E.1}$$

for some constant $c \in \mathbb{R}$, where $\mathbf{1}$ is the all-one vector.

Lemma 35. *For any zero-sum polymatrix game $\mathcal{G}$, it holds that for $\pi, \pi' \in \Delta(S)$ that*

$$\sum_{i \in V} \left[u_i(\pi_i, \pi'_{-i}) + u_i(\pi'_i, \pi_{-i})\right] = 0. \tag{E.2}$$

Or equivalently, $\sum_{i \in V} \left[\pi_i^\top A_i \pi' + (\pi'_i)^\top A_i \pi\right] = 0$. It follows that

$$\sum_{i \in V} \left\langle \pi_i - \pi'_i, A_i(\pi - \pi') \right\rangle = \sum_{i \in V} \left[u_i(\pi) + u_i(\pi')\right] - \sum_{i \in V} \left[\pi_i^\top A_i \pi' + (\pi'_i)^\top A_i \pi\right] = 0.$$

E.1.1 Proof of Theorem 8

We start with the following lemma that characterizes the iterates of OMWU, which generalizes Lemma 13 for zero-sum two-player games to zero-sum polymatrix games. The proof can be found in Appendix E.3.2.

Lemma 36. *The iterates of OMWU based on the update rule (6.12) satisfy*

$$\left\langle \log \pi^{(t+1)} - (1 - \eta\tau) \log \pi^{(t)} - \eta\tau \log \pi_\tau^\star,\, \overline{\pi}^{(t+1)} - \pi_\tau^\star \right\rangle = 0.$$

To continue, by the definition of KL divergence, we have

$$\begin{aligned}
&\left\langle \log \pi^{(t+1)} - (1 - \eta\tau) \log \pi^{(t)} - \eta\tau \log \pi_\tau^\star,\, \overline{\pi}^{(t+1)} \right\rangle \\
&= \left\langle \log \overline{\pi}^{(t+1)} - (1 - \eta\tau) \log \pi^{(t)} - \eta\tau \log \pi_\tau^\star,\, \overline{\pi}^{(t+1)} \right\rangle \\
&\quad - \left\langle \log \overline{\pi}^{(t+1)} - \log \pi^{(t+1)}, \pi^{(t+1)} \right\rangle - \left\langle \log \overline{\pi}^{(t+1)} - \log \pi^{(t+1)},\, \overline{\pi}^{(t+1)} - \pi^{(t+1)} \right\rangle \\
&= (1 - \eta\tau) \mathsf{KL}\left(\overline{\pi}^{(t+1)} \,\|\, \pi^{(t)}\right) + \eta\tau \mathsf{KL}\left(\overline{\pi}^{(t+1)} \,\|\, \pi_\tau^\star\right) + \mathsf{KL}\left(\pi^{(t+1)} \,\|\, \overline{\pi}^{(t+1)}\right) \\
&\quad - \left\langle \log \overline{\pi}^{(t+1)} - \log \pi^{(t+1)},\, \overline{\pi}^{(t+1)} - \pi^{(t+1)} \right\rangle.
\end{aligned}$$

In addition,

$$-\left\langle \log \pi^{(t+1)} - (1-\eta\tau)\log \pi^{(t)} - \eta\tau \log \pi_\tau^\star,\, \pi_\tau^\star \right\rangle = \mathsf{KL}\left(\pi_\tau^\star \,\|\, \pi^{(t+1)}\right) - (1-\eta\tau)\mathsf{KL}\left(\pi_\tau^\star \,\|\, \pi^{(t)}\right).$$

Summing up the above two relations, in view of Lemma 36, it holds that

$$\begin{aligned}\mathsf{KL}\left(\pi_\tau^\star \,\|\, \pi^{(t+1)}\right) = (1-\eta\tau)\mathsf{KL}\left(\pi_\tau^\star \,\|\, \pi^{(t)}\right) - (1-\eta\tau)\mathsf{KL}\left(\overline{\pi}^{(t+1)} \,\|\, \pi^{(t)}\right) - \mathsf{KL}\left(\pi^{(t+1)} \,\|\, \overline{\pi}^{(t+1)}\right) \\ + \left\langle \log \overline{\pi}^{(t+1)} - \log \pi^{(t+1)},\, \overline{\pi}^{(t+1)} - \pi^{(t+1)} \right\rangle - \eta\tau\mathsf{KL}\left(\overline{\pi}^{(t+1)} \,\|\, \pi_\tau^\star\right).\end{aligned} \tag{E.3}$$

We now proceed to bound the terms of interest one by one.

Bounding $\mathsf{KL}\left(\pi_\tau^\star \,\|\, \pi^{(t)}\right)$. We aim to control the right-hand-side (RHS) of (E.3). Based on the update rule of $\overline{\pi}_i^{(t+1)}$ in Algorithm 8, we have

$$\begin{aligned}\log \overline{\pi}_i^{(t+1)} - \log \pi_i^{(t+1)} &\overset{\mathbf{1}}{=} \eta A_i(\overline{\pi}^{(t)} - \overline{\pi}^{(t+1)}) \\ &\overset{\mathbf{1}}{=} \eta A_i(\overline{\pi}^{(t)} - \pi^{(t)}) + \eta A_i(\pi^{(t)} - \overline{\pi}^{(t+1)}).\end{aligned} \tag{E.4}$$

It follows that

$$\begin{aligned}&\left\langle \log \overline{\pi}_i^{(t+1)} - \log \pi_i^{(t+1)},\, \overline{\pi}_i^{(t+1)} - \pi_i^{(t+1)} \right\rangle \\ &= \eta \sum_{j\in\mathcal{N}_i} (\overline{\pi}_i^{(t+1)} - \pi_i^{(t+1)})^\top A_{ij}(\overline{\pi}_j^{(t)} - \pi_j^{(t)}) + \eta \sum_{j\in\mathcal{N}_i} (\overline{\pi}_i^{(t+1)} - \pi_i^{(t+1)})^\top A_{ij}(\pi_j^{(t)} - \overline{\pi}_j^{(t+1)}) \\ &\le \eta \sum_{j\in\mathcal{N}_i} \|A_{ij}\|_\infty \left\|\overline{\pi}_i^{(t+1)} - \pi_i^{(t+1)}\right\|_1 \left\|\pi_j^{(t)} - \overline{\pi}_j^{(t)}\right\|_1 + \eta \sum_{j\in\mathcal{N}_i} \|A_{ij}\|_\infty \left\|\overline{\pi}_i^{(t+1)} - \pi_i^{(t+1)}\right\|_1 \left\|\pi_j^{(t)} - \overline{\pi}_j^{(t+1)}\right\|_1 \\ &\le \frac{\eta}{2}\|A\|_\infty \sum_{j\in\mathcal{N}_i} \left(\left\|\overline{\pi}_j^{(t)} - \pi_j^{(t)}\right\|_1^2 + \left\|\overline{\pi}_j^{(t+1)} - \pi_j^{(t)}\right\|_1^2 + 2\left\|\overline{\pi}_i^{(t+1)} - \pi_i^{(t+1)}\right\|_1^2\right) \\ &\le \eta\|A\|_\infty \sum_{j\in\mathcal{N}_i} \left(\mathsf{KL}\left(\pi_j^{(t)} \,\|\, \overline{\pi}_j^{(t)}\right) + \mathsf{KL}\left(\overline{\pi}_j^{(t+1)} \,\|\, \pi_j^{(t)}\right) + 2\mathsf{KL}\left(\pi_i^{(t+1)} \,\|\, \overline{\pi}_i^{(t+1)}\right)\right),\end{aligned} \tag{E.5}$$

where the last line follows from Pinsker's inequality. Summing the inequality over $i \in V$, we get

$$\begin{aligned}&\left\langle \log \overline{\pi}^{(t+1)} - \log \pi^{(t+1)},\, \overline{\pi}^{(t+1)} - \pi^{(t+1)} \right\rangle \\ &\le \eta d_{\mathsf{max}} \|A\|_\infty \left(\mathsf{KL}\left(\pi^{(t)} \,\|\, \overline{\pi}^{(t)}\right) + \mathsf{KL}\left(\overline{\pi}^{(t+1)} \,\|\, \pi^{(t)}\right) + 2\mathsf{KL}\left(\pi^{(t+1)} \,\|\, \overline{\pi}^{(t+1)}\right)\right).\end{aligned}$$

Plugging the above inequality back into (E.3) yields

$$\begin{aligned}\mathsf{KL}\left(\pi_\tau^\star \,\|\, \pi^{(t+1)}\right) \le (1-\eta\tau)\mathsf{KL}\left(\pi_\tau^\star \,\|\, \pi^{(t)}\right) - (1-\eta\tau-\eta d_{\mathsf{max}}\|A\|_\infty)\,\mathsf{KL}\left(\overline{\pi}^{(t+1)} \,\|\, \pi^{(t)}\right) \\ - (1-2\eta d_{\mathsf{max}}\|A\|_\infty)\mathsf{KL}\left(\pi^{(t+1)} \,\|\, \overline{\pi}^{(t+1)}\right) + \eta d_{\mathsf{max}}\|A\|_\infty\,\mathsf{KL}\left(\pi^{(t)} \,\|\, \overline{\pi}^{(t)}\right) \\ - \eta\tau\mathsf{KL}\left(\overline{\pi}^{(t+1)} \,\|\, \pi_\tau^\star\right).\end{aligned} \tag{E.6}$$

With the choice of the learning rate

$$0 < \eta \le \min\left\{\frac{1}{2\tau},\, \frac{1}{4d_{\mathsf{max}}\|A\|_\infty}\right\},$$

it holds that $1 - \eta\tau - \eta d_{\mathsf{max}} \|A\|_\infty > 0$ and

$$\eta d_{\mathsf{max}} \|A\|_\infty \le \frac{1}{4} \le (1-\eta\tau)(1 - 2\eta d_{\mathsf{max}} \|A\|_\infty). \tag{E.7}$$

This allows us to further relax (E.6) by

$$\begin{aligned}
&\mathsf{KL}\left(\pi_\tau^\star \,\|\, \pi^{(t+1)}\right) + (1 - 2\eta d_{\mathsf{max}} \|A\|_\infty)\mathsf{KL}\left(\pi^{(t+1)} \,\|\, \overline{\pi}^{(t+1)}\right) \\
&\le (1-\eta\tau)\mathsf{KL}\left(\pi_\tau^\star \,\|\, \pi^{(t)}\right) + \eta d_{\mathsf{max}} \|A\|_\infty \,\mathsf{KL}\left(\pi^{(t)} \,\|\, \overline{\pi}^{(t)}\right) \\
&\le (1-\eta\tau)\left(\mathsf{KL}\left(\pi_\tau^\star \,\|\, \pi^{(t)}\right) + (1 - 2\eta d_{\mathsf{max}} \|A\|_\infty)\mathsf{KL}\left(\pi^{(t)} \,\|\, \overline{\pi}^{(t)}\right)\right).
\end{aligned}$$

Let us now introduce the potential function of iterates

$$L^{(t)} := \mathsf{KL}\left(\pi_\tau^\star \,\|\, \pi^{(t)}\right) + (1 - 2\eta d_{\mathsf{max}} \|A\|_\infty)\mathsf{KL}\left(\pi^{(t)} \,\|\, \overline{\pi}^{(t)}\right),$$

which allows us to simply the previous inequality as

$$L^{(t+1)} \le (1-\eta\tau)L^{(t)} \le (1-\eta\tau)^{t+1} L^{(0)} = (1-\eta\tau)^{t+1}\mathsf{KL}\left(\pi_\tau^\star \,\|\, \pi^{(0)}\right), \tag{E.8}$$

where the last equality follows from the definition $\overline{\pi}^{(0)} = \pi^{(0)}$. Hence, we have

$$\mathsf{KL}\left(\pi_\tau^\star \,\|\, \pi^{(t)}\right) \le L^{(t)} \le (1-\eta\tau)^t \mathsf{KL}\left(\pi_\tau^\star \,\|\, \pi^{(0)}\right).$$

Bounding $\mathsf{KL}\left(\pi_\tau^\star \,\|\, \overline{\pi}^{(t+1)}\right)$. Following similar approaches to (E.5), we can bound

$$\begin{aligned}
&-\left\langle \pi_{i,\tau}^\star - \overline{\pi}_i^{(t+1)}, \log \overline{\pi}_i^{(t+1)} - \log \pi_i^{(t+1)} \right\rangle \\
&= \eta(\overline{\pi}_i^{(t+1)} - \pi_{i,\tau}^\star)^\top A_i(\overline{\pi}^{(t)} - \pi^{(t)}) + \eta(\overline{\pi}_i^{(t+1)} - \pi_{i,\tau}^\star)^\top A_i(\pi^{(t)} - \overline{\pi}^{(t+1)}) \\
&\le \eta \|A\|_\infty \sum_{j\in\mathcal{N}_i} \left(\mathsf{KL}\left(\pi_j^{(t)} \,\|\, \overline{\pi}_j^{(t)}\right) + \mathsf{KL}\left(\overline{\pi}_j^{(t+1)} \,\|\, \pi_j^{(t)}\right) + 2\mathsf{KL}\left(\pi_{i,\tau}^\star \,\|\, \overline{\pi}_i^{(t+1)}\right)\right).
\end{aligned} \tag{E.9}$$

Summing the inequality over $i \in V$ leads to

$$\begin{aligned}
&-\left\langle \pi_\tau^\star - \overline{\pi}^{(t+1)}, \log \overline{\pi}^{(t+1)} - \log \pi^{(t+1)} \right\rangle \\
&\le \eta d_{\mathsf{max}} \|A\|_\infty \left[\mathsf{KL}\left(\pi^{(t)} \,\|\, \overline{\pi}^{(t)}\right) + \mathsf{KL}\left(\overline{\pi}^{(t+1)} \,\|\, \pi^{(t)}\right) + 2\mathsf{KL}\left(\pi_\tau^\star \,\|\, \overline{\pi}^{(t+1)}\right)\right].
\end{aligned}$$

On the other hand, by the definition of KL divergence, we have

$$\mathsf{KL}\left(\pi_\tau^\star \,\|\, \overline{\pi}^{(t+1)}\right) = \mathsf{KL}\left(\pi_\tau^\star \,\|\, \pi^{(t+1)}\right) - \mathsf{KL}\left(\overline{\pi}^{(t+1)} \,\|\, \pi^{(t+1)}\right) - \left\langle \pi_\tau^\star - \overline{\pi}^{(t+1)}, \log \overline{\pi}^{(t+1)} - \log \pi^{(t+1)} \right\rangle. \tag{E.10}$$

Combining the above two inequalities, we get

$$\begin{aligned}
&(1 - 2\eta d_{\mathsf{max}} \|A\|_\infty)\mathsf{KL}\left(\pi_\tau^\star \,\|\, \overline{\pi}^{(t+1)}\right) \\
&\le \mathsf{KL}\left(\pi_\tau^\star \,\|\, \pi^{(t+1)}\right) + \eta d_{\mathsf{max}} \|A\|_\infty \left(\mathsf{KL}\left(\pi^{(t)} \,\|\, \overline{\pi}^{(t)}\right) + \mathsf{KL}\left(\overline{\pi}^{(t+1)} \,\|\, \pi^{(t)}\right)\right).
\end{aligned}$$

Plugging the above inequality back into (E.6), we have

$$
\begin{aligned}
&(1-2\eta d_{\mathsf{max}}\left\|A\right\|_{\infty})\mathsf{KL}\left(\pi_{\tau}^{\star}\left\|\,\overline{\pi}^{(t+1)}\right.\right)\\
&\leq(1-\eta\tau)\mathsf{KL}\left(\pi_{\tau}^{\star}\left\|\,\pi^{(t)}\right.\right)-(1-\eta\tau-2d_{\mathsf{max}}\eta\left\|A\right\|_{\infty})\mathsf{KL}\left(\overline{\pi}^{(t+1)}\left\|\,\pi^{(t)}\right.\right)-\eta\tau\mathsf{KL}\left(\overline{\pi}^{(t+1)}\left\|\,\pi_{\tau}^{\star}\right.\right)\\
&\quad-(1-2\eta d_{\mathsf{max}}\left\|A\right\|_{\infty})\mathsf{KL}\left(\pi^{(t+1)}\left\|\,\overline{\pi}^{(t+1)}\right.\right)+2\eta d_{\mathsf{max}}\left\|A\right\|_{\infty}\mathsf{KL}\left(\pi^{(t)}\left\|\,\overline{\pi}^{(t)}\right.\right)\\
&\leq(1-\eta\tau)\mathsf{KL}\left(\pi_{\tau}^{\star}\left\|\,\pi^{(t)}\right.\right)+2\eta d_{\mathsf{max}}\left\|A\right\|_{\infty}\mathsf{KL}\left(\pi^{(t)}\left\|\,\overline{\pi}^{(t)}\right.\right)\\
&\leq\mathsf{KL}\left(\pi_{\tau}^{\star}\left\|\,\pi^{(t)}\right.\right)+(1-2\eta d_{\mathsf{max}}\left\|A\right\|_{\infty})\mathsf{KL}\left(\pi^{(t)}\left\|\,\overline{\pi}^{(t)}\right.\right)=L^{(t)},
\end{aligned}
$$

where the second and third inequalities follow from the choice of the learning rate, and the last line follows from the definition of the potential function $L^{(t)}$. Then the result follows from (E.8) as

$$
\frac{1}{2}\mathsf{KL}\left(\pi_{\tau}^{\star}\left\|\,\overline{\pi}^{(t+1)}\right.\right)\leq(1-2\eta d_{\mathsf{max}}\left\|A\right\|_{\infty})\mathsf{KL}\left(\pi_{\tau}^{\star}\left\|\,\overline{\pi}^{(t+1)}\right.\right)\leq L^{(t)}\leq(1-\eta\tau)^{t}\mathsf{KL}\left(\pi_{\tau}^{\star}\left\|\,\pi^{(0)}\right.\right).
$$

Bounding the QRE-Gap. Finally, we bound the QRE-gap, which can be linked to the KL divergence using the following lemma. The proof can be found in Appendix E.3.3.

Lemma 37. *For any $\pi\in\Delta(S)$ and QRE $\pi_{\tau}^{\star}\in\Delta(S)$, it holds that*

$$
\texttt{QRE-Gap}_{\tau}(\pi)\leq\tau\mathsf{KL}\left(\pi\left\|\,\pi_{\tau}^{\star}\right.\right)+\frac{d_{\mathsf{max}}^{2}\left\|A\right\|_{\infty}^{2}}{\tau}\mathsf{KL}\left(\pi_{\tau}^{\star}\left\|\,\pi\right.\right).
$$

Lemma 37 tells us

$$
\texttt{QRE-Gap}_{\tau}(\overline{\pi}^{(t)})\leq\tau\mathsf{KL}\left(\overline{\pi}^{(t)}\left\|\,\pi_{\tau}^{\star}\right.\right)+\frac{d_{\mathsf{max}}^{2}\left\|A\right\|_{\infty}^{2}}{\tau}\mathsf{KL}\left(\pi_{\tau}^{\star}\left\|\,\overline{\pi}^{(t)}\right.\right).\tag{E.11}
$$

With $\mathsf{KL}\left(\pi_{\tau}^{\star}\left\|\,\overline{\pi}^{(t)}\right.\right)$ controlled in the above, we still need to control $\mathsf{KL}\left(\overline{\pi}^{(t)}\left\|\,\pi_{\tau}^{\star}\right.\right)$. From (E.6), it follows that

$$
\tau\mathsf{KL}\left(\overline{\pi}^{(t)}\left\|\,\pi_{\tau}^{\star}\right.\right)\leq\eta^{-1}(1-\eta\tau)L^{(t-1)}\leq\eta^{-1}(1-\eta\tau)^{t}L^{(0)}=\eta^{-1}(1-\eta\tau)^{t}\mathsf{KL}\left(\pi_{\tau}^{\star}\left\|\,\pi^{(0)}\right.\right).
$$

Plugging them back to (E.11), we arrive at

$$
\texttt{QRE-Gap}_{\tau}(\overline{\pi}^{(t)})\leq\left(\eta^{-1}+2\tau^{-1}d_{\mathsf{max}}^{2}\|A\|_{\infty}^{2}\right)(1-\eta\tau)^{t-1}\mathsf{KL}\left(\pi_{\tau}^{\star}\left\|\,\pi^{(0)}\right.\right).
$$

E.1.2 Proof of Theorem 9

We begin with bounding the KL divergence $\mathsf{KL}\left(\pi_{\tau}^{\star}\left\|\,\pi^{(t)}\right.\right)$ and then move to bound the QRE-gap by linking it to the KL divergence.

Bounding the term $\mathsf{KL}\left(\pi_{\tau}^{\star}\left\|\,\pi^{(t)}\right.\right)$. We start with the following equation

$$
\begin{aligned}
(1-\eta\tau)\mathsf{KL}\left(\pi_{i,\tau}^{\star}\left\|\,\pi_{i}^{(t)}\right.\right)&=(1-\eta\tau)\mathsf{KL}\left(\overline{\pi}_{i}^{(t+1)}\left\|\,\pi_{i}^{(t)}\right.\right)+\eta\tau\mathsf{KL}\left(\overline{\pi}_{i}^{(t+1)}\left\|\,\pi_{i,\tau}^{\star}\right.\right)+\mathsf{KL}\left(\pi_{i}^{(t+1)}\left\|\,\overline{\pi}_{i}^{(t+1)}\right.\right)\\
&\quad+\mathsf{KL}\left(\pi_{i,\tau}^{\star}\left\|\,\pi_{i}^{(t+1)}\right.\right)-\left\langle\log\overline{\pi}_{i}^{(t+1)}-\log\pi_{i}^{(t+1)},\,\overline{\pi}_{i}^{(t+1)}-\pi_{i}^{(t+1)}\right\rangle\\
&\quad+\eta(\overline{\pi}_{i}^{(t+1)}-\pi_{i,\tau}^{\star})^{\top}A_{i}(\overline{\pi}^{(\kappa_{i}^{(t)})}-\pi_{\tau}^{\star})
\end{aligned}\tag{E.12}
$$

where its proof follows a similar deduction as (E.3). Our first target is to bound the last two terms on the RHS of (E.12) with

$$\eta\tau \mathsf{KL}\left(\overline{\pi}_i^{(t+1)} \,\|\, \pi_{i,\tau}^{\star}\right) + \mathsf{KL}\left(\pi_i^{(t+1)} \,\|\, \overline{\pi}_i^{(t+1)}\right) + (1-\eta\tau)\mathsf{KL}\left(\overline{\pi}_i^{(t+1)} \,\|\, \pi_i^{(t)}\right).$$

Let us introduce the potential function of iterates

$$\Psi_i^{(l)} := \mathsf{KL}\left(\overline{\pi}_i^{(l+1)} \,\|\, \pi_i^{(l)}\right) + \mathsf{KL}\left(\pi_i^{(l)} \,\|\, \overline{\pi}_i^{(l)}\right), \qquad \Psi^{(l)} = \sum_{i\in V} \Psi_i^{(l)} = \mathsf{KL}\left(\overline{\pi}^{(l+1)} \,\|\, \pi^{(l)}\right) + \mathsf{KL}\left(\pi^{(l)} \,\|\, \overline{\pi}^{(l)}\right),$$

which will be used repetitively in the rest of this proof. For notational simplicity, let $\Psi_i^{(l)} = 0$ when $l < 0$.

Step 1: bounding $\left\langle \log \overline{\pi}_i^{(t+1)} - \log \pi_i^{(t+1)}, \overline{\pi}_i^{(t+1)} - \pi_i^{(t+1)} \right\rangle$**.** Following a similar argument as (E.5), we get

$$\begin{aligned}
&\left\langle \log \overline{\pi}_i^{(t+1)} - \log \pi_i^{(t+1)}, \overline{\pi}_i^{(t+1)} - \pi_i^{(t+1)} \right\rangle \\
&= \eta \sum_{j\in\mathcal{N}_i} (\overline{\pi}_i^{(t+1)} - \pi_i^{(t+1)})^\top A_{ij} (\overline{\pi}_j^{(\kappa_i^{(t+1)})} - \overline{\pi}_j^{(\kappa_i^{(t)})}) \\
&\le \eta d_{\mathsf{max}} \|A\|_\infty \mathsf{KL}\left(\pi_i^{(t+1)} \,\|\, \overline{\pi}_i^{(t+1)}\right) + \frac{\eta \|A\|_\infty}{2} \sum_{j\in\mathcal{N}_i} \|\overline{\pi}_j^{(\kappa_i^{(t+1)})} - \overline{\pi}_j^{(\kappa_i^{(t)})}\|_1^2.
\end{aligned} \tag{E.13}$$

To control the term $\|\overline{\pi}_j^{(\kappa_i^{(t+1)})} - \overline{\pi}_j^{(\kappa_i^{(t)})}\|_1^2$, when $t = 0$, we have

$$\|\overline{\pi}_j^{(\kappa_i^{(t+1)})} - \overline{\pi}_j^{(\kappa_i^{(t)})}\|_1^2 = \|\overline{\pi}_j^{(\kappa_i^{(t+1)})} - \overline{\pi}_j^{(0)}\|_1^2 \le \|\overline{\pi}_j^{(1)} - \pi_j^{(0)}\|_1^2 \le 2\Psi_j^{(0)} \tag{E.14}$$

by Pinsker's inequality. For $t \ge 1$, consider the decomposition

$$\overline{\pi}_j^{(t)} - \overline{\pi}_j^{(t-k)} = \sum_{l=t-k}^{t-1} \left(\overline{\pi}_j^{(l+1)} - \overline{\pi}_j^{(l)}\right), \quad \forall 1 \le k \le t,$$

it then follows that

$$\begin{aligned}
\|\overline{\pi}_j^{(t)} - \overline{\pi}_j^{(t-k)}\|_1^2 &\le k \sum_{l=t-k}^{t-1} \|\overline{\pi}_j^{(l+1)} - \overline{\pi}_j^{(l)}\|_1^2 \\
&\le 2k \sum_{l=t-k}^{t-1} \left(\|\overline{\pi}_j^{(l+1)} - \pi_j^{(l)}\|_1^2 + \|\pi_j^{(l)} - \overline{\pi}_j^{(l)}\|_1^2\right) \\
&\le 4k \sum_{l=t-k}^{t-1} \Psi_j^{(l)},
\end{aligned} \tag{E.15}$$

where the last line applies Pinsker's inequality. Depending on whether $\gamma_i^{(t+1)} > 0$, we proceed to bound the terms $\|\overline{\pi}_j^{(\kappa_i^{(t+1)})} - \overline{\pi}_j^{(\kappa_i^{(t)})}\|_1^2$ in (E.13) considering the following two cases based on (E.15).

- $\gamma_i^{(t+1)} = 0$. Then

$$\begin{aligned}\big\|\overline{\pi}_j^{(\kappa_i^{(t+1)})} - \overline{\pi}_j^{(\kappa_i^{(t)})}\big\|_1^2 &\le 2\big\|\overline{\pi}_j^{(t+1)} - \overline{\pi}_j^{(t)}\big\|_1^2 + 2\big\|\overline{\pi}_j^{(t)} - \overline{\pi}_j^{(\kappa_i^{(t)})}\big\|_1^2 \\ &\le 8\Psi_j^{(t)} + 8\gamma_i^{(t)} \sum_{l=t-\gamma_i^{(t)}}^{t-1} \Psi_j^{(l)},\end{aligned}$$

  where the last step uses (E.15) and

$$\big\|\overline{\pi}_j^{(t+1)} - \overline{\pi}_j^{(t)}\big\|_1^2 \le 2\left(\big\|\overline{\pi}_j^{(t+1)} - \pi_j^{(t)}\big\|_1^2 + \big\|\pi_j^{(t)} - \overline{\pi}_j^{(t)}\big\|_1^2\right) \le 4\Psi_j^{(t)}$$

  via again Pinsker's inequality.

- $\gamma_i^{(t+1)} > 0$. Then it follows similarly that

$$\begin{aligned}\big\|\overline{\pi}_j^{(\kappa_i^{(t+1)})} - \overline{\pi}_j^{(\kappa_i^{(t)})}\big\|_1^2 &\le \sum_{l=t+1-\gamma_i^{(t+1)}}^{t-1} \big\|\overline{\pi}_j^{(l+1)} - \overline{\pi}_j^{(l)}\big\|_1^2 + \sum_{l=t-\gamma_i^{(t)}}^{t-1} \big\|\overline{\pi}_j^{(l+1)} - \overline{\pi}_j^{(l)}\big\|_1^2 \\ &\le 4\gamma_i^{(t+1)} \sum_{l=t-\gamma_i^{(t+1)}}^{t-1} \Psi_j^{(l)} + 4\gamma_i^{(t)} \sum_{l=t-\gamma_i^{(t)}}^{t-1} \Psi_j^{(l)}.\end{aligned}$$

Combining the above two bounds together, we get

$$\big\|\overline{\pi}_j^{(\kappa_i^{(t+1)})} - \overline{\pi}_j^{(\kappa_i^{(t)})}\big\|_1^2 \le 8\Psi_j^{(t)} + 8\gamma_i^{(t)} \sum_{l=t-\gamma_i^{(t)}}^{t-1} \Psi_j^{(l)} + 4\gamma_i^{(t+1)} \sum_{l=t-\gamma_i^{(t+1)}}^{t-1} \Psi_j^{(l)} \tag{E.16}$$

when $t > 0$. In view of (E.14) when $t = 0$, the above bound (E.16) holds for all $t \ge 0$. Plugging the above inequality into (E.13) yields

$$\begin{aligned}\left\langle \log \overline{\pi}_i^{(t+1)} - \log \pi_i^{(t+1)},\, \overline{\pi}_i^{(t+1)} - \pi_i^{(t+1)} \right\rangle &\le 2\eta \,\|A\|_\infty \sum_{j\in\mathcal{N}_i} \sum_{l=t-\gamma_i^{(t+1)}}^{t-1} \gamma_i^{(t+1)} \Psi_j^{(l)} + 4\eta \,\|A\|_\infty \sum_{j\in\mathcal{N}_i} \sum_{l=t-\gamma_i^{(t)}}^{t-1} \gamma_i^{(t)} \Psi_j^{(l)} \\ &\quad + 4\eta \,\|A\|_\infty \sum_{j\in\mathcal{N}_i} \Psi_j^{(t)} + \eta d_{\mathsf{max}} \,\|A\|_\infty \,\mathsf{KL}\left(\pi_i^{(t+1)} \,\|\, \overline{\pi}_i^{(t+1)}\right).\end{aligned} \tag{E.17}$$

Step 2: bounding $(\overline{\pi}_i^{(t+1)} - \pi_{i,\tau}^\star)^\top A_i(\overline{\pi}^{(\kappa_i^{(t+1)})} - \pi_\tau^\star)$**.** Let us begin with the following decomposition

$$\begin{aligned}(\overline{\pi}_i^{(t+1)} - \pi_{i,\tau}^\star)^\top A_i(\overline{\pi}^{(\kappa_i^{(t+1)})} - \pi_\tau^\star) &= (\overline{\pi}_i^{(t+1)} - \pi_{i,\tau}^\star)^\top A_i(\overline{\pi}^{(t+1)} - \pi_\tau^\star) \\ &\quad + (\overline{\pi}_i^{(t+1)} - \pi_{i,\tau}^\star)^\top A_i(\overline{\pi}^{(\kappa_i^{(t+1)})} - \overline{\pi}^{(t+1)}),\end{aligned} \tag{E.18}$$

where the second term in the RHS of (E.18) can be bounded by

$$
\begin{aligned}
&(\overline{\pi}_i^{(t+1)} - \pi_{i,\tau}^\star)^\top A_i(\overline{\pi}^{(\kappa_i^{(t+1)})} - \overline{\pi}^{(t+1)}) \\
&= \sum_{j\in\mathcal{N}_i} (\overline{\pi}_i^{(t+1)} - \pi_{i,\tau}^\star)^\top A_{ij}(\overline{\pi}_j^{(\kappa_i^{(t+1)})} - \overline{\pi}_j^{(t+1)}) \\
&\le \|A\|_\infty \sum_{j\in\mathcal{N}_i} \big\|\overline{\pi}_i^{(t+1)} - \pi_{i,\tau}^\star\big\|_1 \big\|\overline{\pi}_j^{(\kappa_i^{(t+1)})} - \overline{\pi}_j^{(t+1)}\big\|_1 \\
&\le \frac{1}{2}\|A\|_\infty \sum_{j\in\mathcal{N}_i} \left(\frac{\tau}{d_{\max}\|A\|_\infty}\big\|\overline{\pi}_i^{(t+1)} - \pi_{i,\tau}^\star\big\|_1^2 + \frac{d_{\max}\|A\|_\infty}{\tau}\big\|\overline{\pi}_j^{(\kappa_i^{(t+1)})} - \overline{\pi}_j^{(t+1)}\big\|_1^2\right) \\
&\le \tau \mathsf{KL}\left(\overline{\pi}_i^{(t+1)} \,\|\, \pi_{i,\tau}^\star\right) + \frac{d_{\max}\|A\|_\infty^2}{2\tau} \sum_{j\in\mathcal{N}_i} \big\|\overline{\pi}_j^{(\kappa_i^{(t+1)})} - \overline{\pi}_j^{(t+1)}\big\|_1^2.
\end{aligned}
$$

Following similar deduction of (E.16) for the second term, we attain

$$
\begin{aligned}
&(\overline{\pi}_i^{(t+1)} - \pi_{i,\tau}^\star)^\top A_i(\overline{\pi}^{(\kappa_i^{(t+1)})} - \overline{\pi}^{(t+1)}) \\
&\le \tau \mathsf{KL}\left(\overline{\pi}_i^{(t+1)} \,\|\, \pi_{i,\tau}^\star\right) + \frac{4d_{\max}\|A\|_\infty^2}{\tau} \sum_{j\in\mathcal{N}_i} \left(\Psi_j^{(t)} + \sum_{l=t-\gamma_i^{(t+1)}}^{t-1} \gamma_i^{(t+1)} \Psi_j^{(l)}\right).
\end{aligned}
$$

Plugging the above inequality back to (E.18) results in

$$
\begin{aligned}
(\overline{\pi}_i^{(t+1)} - \pi_{i,\tau}^\star)^\top A_i(\overline{\pi}^{(\kappa_i^{(t+1)})} - \pi_\tau^\star) &\le (\overline{\pi}_i^{(t+1)} - \pi_{i,\tau}^\star)^\top A_i(\overline{\pi}^{(t+1)} - \pi_\tau^\star) \\
&\quad + \tau \mathsf{KL}\left(\overline{\pi}_i^{(t+1)} \,\|\, \pi_{i,\tau}^\star\right) + \frac{4d_{\max}\|A\|_\infty^2}{\tau} \sum_{j\in\mathcal{N}_i} \left(\Psi_j^{(t)} + \sum_{l=t-\gamma_i^{(t+1)}}^{t-1} \gamma_i^{(t+1)} \Psi_j^{(l)}\right). \qquad \text{(E.19)}
\end{aligned}
$$

Step 3: combining the bounds. For simplicity, we introduce the short-hand notation

$$
c_\tau = 1 + \frac{d_{\max}\|A\|_\infty}{\tau} \qquad \text{and} \qquad c_A = d_{\max}\|A\|_\infty. \qquad \text{(E.20)}
$$

Combining (E.17) and (E.19) into (E.12), and summing over $i \in V$ gives

$$
\begin{aligned}
&(1-\eta\tau)\mathsf{KL}\left(\pi_\tau^\star \,\|\, \pi^{(t)}\right) \\
&\ge (1-\eta\tau)\mathsf{KL}\left(\overline{\pi}^{(t+1)} \,\|\, \pi^{(t)}\right) + (1-2\eta c_A)\mathsf{KL}\left(\pi^{(t+1)} \,\|\, \overline{\pi}^{(t+1)}\right) + \mathsf{KL}\left(\pi_\tau^\star \,\|\, \pi^{(t+1)}\right) \\
&\quad - 4\eta\|A\|_\infty \sum_{i\in V}\sum_{j\in\mathcal{N}_i} \left(\sum_{l=t-\gamma_i^{(t+1)}}^{t-1} c_\tau \gamma_i^{(t+1)} \Psi_j^{(l)} + \sum_{l=t-\gamma_i^{(t)}}^{t-1} \gamma_i^{(t)} \Psi_j^{(l)} + c_\tau \Psi_j^{(t)}\right) \\
&\ge \mathsf{KL}\left(\pi_\tau^\star \,\|\, \pi^{(t+1)}\right) + (1 - 4\eta c_A(c_\tau+1))\,\Psi^{(t)} \\
&\quad - 4\eta\|A\|_\infty \sum_{i\in V}\sum_{j\in\mathcal{N}_i} \left(c_\tau \sum_{l=t-\gamma_i^{(t+1)}}^{t-1} \gamma_i^{(t+1)} \Psi_j^{(l)} + \sum_{l=t-\gamma_i^{(t)}}^{t-1} \gamma_i^{(t)} \Psi_j^{(l)}\right), \qquad \text{(E.21)}
\end{aligned}
$$

where we make use of the fact

$$
\sum_{i\in V} (\overline{\pi}_i^{(t+1)} - \pi_{i,\tau}^\star)^\top A_i(\overline{\pi}^{(t+1)} - \pi_\tau^\star) = 0
$$

from Lemma 35 in the first inequality, and the second inequality uses the relation

$$\sum_{i\in V}\sum_{j\in\mathcal{N}_i}\Psi_j^{(t)}=\sum_{i\in V}d_i\Psi_i^{(t)}\le d_{\mathsf{max}}\Psi^{(t)}.$$

Step 4: finishing up via averaging the delay. We now evaluate the expectation of $\mathsf{KL}\left(\pi_\tau^\star\,\|\,\pi^{(t+1)}\right)$. Recall that we use subscript $\mathbb{E}_{s_t}[\cdot]$ to represent the conditional expectation given $s_t=\{\gamma_i^{(t)}\}_{i\in V}$. We shall first control the conditional expectation of the last term in (E.21). Observing that $\overline{\pi}_j^{(l+1)},\pi_j^{(l)}$ are independent of $\gamma_i^{(t)}$ for $j\in\mathcal{N}_i$ and $l\le t-1$. Using the definition of $E(t-l)$, we have

$$\begin{aligned}\sum_{i\in V}\sum_{j\in\mathcal{N}_i}\mathbb{E}_{s_t}\left[\gamma_i^{(t)}\sum_{l=t-\gamma_i^{(t)}}^{t-1}\Psi_j^{(l)}\right]&=\sum_{i\in V}\sum_{l=0}^{t-1}\sum_{j\in\mathcal{N}_i}\mathbb{E}_{t-l\le\gamma_i^{(t)}}\left[\gamma_i^{(t)}\Psi_j^{(l)}\right]\\&\le\sum_{l=0}^{t-1}E(t-l)\sum_{i\in V}\sum_{j\in\mathcal{N}_i}\Psi_j^{(l)}\\&=\sum_{l=0}^{t-1}E(t-l)\sum_{i\in V}\sum_{j\in\mathcal{N}_i}\Psi_i^{(l)}\\&\le d_{\mathsf{max}}\sum_{l=0}^{t-1}E(t-l)\sum_{i\in V}\Psi_i^{(l)}=d_{\mathsf{max}}\sum_{l=0}^{t-1}E(t-l)\Psi^{(l)},\end{aligned}\tag{E.22}$$

where the second line follows from the definition of $E(t-l)$ in Assumption 4. Applying a similar argument to bound $\sum_{i\in V}\sum_{j\in\mathcal{N}_i}\mathbb{E}_{\gamma^{(t+1)}}\left[\gamma_i^{(t+1)}\sum_{l=t-\gamma_i^{(t+1)}}^{t-1}\Psi_j^{(l)}\right]$, and taking expectation of $s_t,\gamma^{(t+1)}$ on both sides of (E.21), we get

$$\begin{aligned}(1-\eta\tau)\mathbb{E}_{s_t}\left[\mathsf{KL}\left(\pi_\tau^\star\,\|\,\pi^{(t)}\right)\right]&\ge\mathbb{E}_{s_t,\gamma^{(t+1)}}\left[\mathsf{KL}\left(\pi_\tau^\star\,\|\,\pi^{(t+1)}\right)+(1-4c_A(c_\tau+1))\Psi^{(t)}\right]\\&\quad-4\eta c_A(c_\tau+1)\sum_{l=0}^{t-1}E(t-l)\Psi^{(l)}.\end{aligned}$$

Taking expectation on both sides over all the delays yields

$$\begin{aligned}&(1-\eta\tau)\mathbb{E}\left[\mathsf{KL}\left(\pi_\tau^\star\,\|\,\pi^{(t)}\right)\right]\\&\ge\mathbb{E}\left[\mathsf{KL}\left(\pi_\tau^\star\,\|\,\pi^{(t+1)}\right)\right]+\mathbb{E}\Big[\underbrace{(1-4\eta c_A(c_\tau+1))\,\Psi^{(t)}-4\eta c_A(c_\tau+1)\textstyle\sum_{l=0}^{t-1}E(t-l)\Psi^{(l)}}_{=:U^{(t)}}\Big].\end{aligned}\tag{E.23}$$

Telescoping over $t=0,1,\ldots,T$, we get

$$(1-\eta\tau)^{T+1}\mathsf{KL}\left(\pi_\tau^\star\,\|\,\pi^{(0)}\right)\ge\mathbb{E}\left[\mathsf{KL}\left(\pi_\tau^\star\,\|\,\pi^{(T+1)}\right)\right]+\sum_{t=0}^{T}(1-\eta\tau)^{T-t}\mathbb{E}\left[U^{(t)}\right],\tag{E.24}$$

which leads to the desired bound if

$$\sum_{t=0}^{t}(1-\eta\tau)^{T-t}\mathbb{E}\left[U^{(t)}\right]\ge 0.\tag{E.25}$$

Proof of (E.25). To begin, notice that with the choice of the learning rate

$$0 < \eta \leq \min\left\{\frac{\tau}{24d_{\mathsf{max}}^2 \left\|A\right\|_\infty^2 (L+1)}, \frac{\zeta-1}{\zeta\tau}\right\},$$

it follows that

$$\frac{1}{1-\eta\tau} \leq \zeta \tag{E.26a}$$

and

$$\begin{aligned} 4\eta c_A(c_\tau+1)(L+1) &< 4\frac{\tau}{24d_{\mathsf{max}}^2 \left\|A\right\|_\infty^2 (L+1)} d_{\mathsf{max}} \left\|A\right\|_\infty \left(2+\frac{d_{\mathsf{max}}\left\|A\right\|_\infty}{\tau}\right)(L+1) \\ &= \frac{\tau}{6d_{\mathsf{max}}\left\|A\right\|_\infty}\left(2+\frac{d_{\mathsf{max}}\left\|A\right\|_\infty}{\tau}\right) = \frac{\tau}{3d_{\mathsf{max}}\left\|A\right\|_\infty} + \frac{1}{6} \leq \frac{1}{2} \end{aligned} \tag{E.26b}$$

as $\tau \leq d_{\mathsf{max}}\left\|A\right\|_\infty$. Both of these relations will be useful in our follow-up analysis.

Now, taking the definition of $U^{(t)}$ (cf. (E.23)), we have

$$\sum_{t=0}^{T}(1-\eta\tau)^{T-t}U^{(t)} = \sum_{t=0}^{T}(1-\eta\tau)^{T-t}\left[\left(1-4\eta c_A(c_\tau+1)\right)\Psi^{(t)} - 4\eta c_A(c_\tau+1)\sum_{l=0}^{t-1}E(t-l)\Psi^{(l)}\right],$$

where the second half of the RHS can be further controlled via

$$\begin{aligned} \sum_{t=0}^{T}(1-\eta\tau)^{T-t}\sum_{l=0}^{t-1}E(t-l)\Psi^{(l)} &= \sum_{t=0}^{T}\Psi^{(t)}\sum_{l=t+1}^{T}(1-\eta\tau)^{T-l}E(l-t) \\ &\leq \sum_{t=0}^{T}\Psi^{(t)}\sum_{l'=0}^{T-t}(1-\eta\tau)^{T-(t+l')}E(l') \\ &= \sum_{t=0}^{T}(1-\eta\tau)^{T-t}\Psi^{(t)}\sum_{l'=0}^{T-t}(1-\eta\tau)^{-l'}E(l') \\ &\leq \sum_{t=0}^{T}(1-\eta\tau)^{T-t}\Psi^{(t)}\sum_{l=0}^{\infty}\zeta^l E(l) \\ &= \sum_{t=0}^{T}(1-\eta\tau)^{T-t}L\Psi^{(t)}, \end{aligned}$$

where the first line follows by changing the order of summation, the second line follows from the change of variable $l' = l - t$, and the last line follows from (E.26a) and the definition of L in Assumption 4. Plugging the above relation back leads to

$$\begin{aligned} \sum_{t=0}^{T}(1-\eta\tau)^{T-t}U^{(t)} &\geq \sum_{t=0}^{T}(1-\eta\tau)^{T-t}\left[\left(1-4\eta c_A(c_\tau+1)\right) - 4\eta c_A(c_\tau+1)L\right]\Psi^{(t)} \\ &\geq \sum_{t=0}^{T}\frac{1}{2}(1-\eta\tau)^{T-t}\Psi^{(t)} \geq 0, \end{aligned} \tag{E.27}$$

where the second line results from (E.26b).

Bounding the term $\mathsf{KL}\left(\pi_\tau^\star \,\|\, \overline{\pi}^{(t+1)}\right)$. With a similar deduction of (E.3), we get

$$(1-\eta\tau)\mathsf{KL}\left(\pi_\tau^\star \,\|\, \pi^{(t)}\right) + \eta\sum_{i\in V}(\overline{\pi}_i^{(t+1)} - \pi_\tau^\star)^\top A_i(\overline{\pi}^{(\kappa_i^{(t)})} - \pi_\tau^\star) = \mathsf{KL}\left(\pi_\tau^\star \,\|\, \overline{\pi}^{(t+1)}\right) + (1-\eta\tau)\mathsf{KL}\left(\overline{\pi}^{(t+1)} \,\|\, \pi^{(t)}\right) + \eta\tau\mathsf{KL}\left(\overline{\pi}^{(t+1)} \,\|\, \pi_\tau^\star\right). \tag{E.28}$$

Following the similar argument of (E.19), we have

$$\begin{aligned}(\overline{\pi}_i^{(t+1)} - \pi_{i,\tau}^\star)^\top A_i(\overline{\pi}^{(\kappa_i^{(t)})} - \pi_\tau^\star) &\le (\overline{\pi}_i^{(t+1)} - \pi_{i,\tau}^\star)^\top A_i(\overline{\pi}^{(t+1)} - \pi_\tau^\star) \\ &\quad + \frac{\tau}{2}\mathsf{KL}\left(\overline{\pi}_i^{(t+1)} \,\|\, \pi_{i,\tau}^\star\right) + \frac{8d_{\mathsf{max}}\|A\|_\infty^2}{\tau}\sum_{j\in\mathcal{N}_i}\left(\Psi_j^{(t)} + \sum_{l=t-\gamma_i^{(t)}}^{t-1}\gamma_i^{(t)}\Psi_j^{(l)}\right).\end{aligned}$$

Summing over $i \in V$ and plugging into (E.28) yields

$$\begin{aligned}&(1-\eta\tau)\mathsf{KL}\left(\pi_\tau^\star \,\|\, \pi^{(t)}\right) + \frac{8\eta d_{\mathsf{max}}\|A\|_\infty^2}{\tau}\sum_{(i,j)\in E}\left(\Psi_j^{(t)} + \sum_{l=t-\gamma_i^{(t)}}^{t-1}\gamma_i^{(t)}\Psi_j^{(l)}\right) \\ &\ge \mathsf{KL}\left(\pi_\tau^\star \,\|\, \overline{\pi}^{(t+1)}\right) + (1-\eta\tau)\mathsf{KL}\left(\overline{\pi}^{(t+1)} \,\|\, \pi^{(t)}\right) + \frac{\eta\tau}{2}\mathsf{KL}\left(\overline{\pi}^{(t+1)} \,\|\, \pi_\tau^\star\right) \\ &\ge \mathsf{KL}\left(\pi_\tau^\star \,\|\, \overline{\pi}^{(t+1)}\right) + \frac{\eta\tau}{2}\mathsf{KL}\left(\overline{\pi}^{(t+1)} \,\|\, \pi_\tau^\star\right).\end{aligned}$$

Taking expectation on both sides over all delays and using (E.22) leads to

$$\begin{aligned}&(1-\eta\tau)\mathbb{E}\left[\mathsf{KL}\left(\pi_\tau^\star \,\|\, \pi^{(t)}\right)\right] + \frac{8\eta d_{\mathsf{max}}^2\|A\|_\infty^2}{\tau}\mathbb{E}\left[\Psi^{(t)} + \sum_{l=0}^{t-1}E(t-l)\Psi^{(l)}\right] \\ &\ge \mathbb{E}\left[\mathsf{KL}\left(\pi_\tau^\star \,\|\, \overline{\pi}^{(t+1)}\right)\right] + \frac{\eta\tau}{2}\mathbb{E}\left[\mathsf{KL}\left(\overline{\pi}^{(t+1)} \,\|\, \pi_\tau^\star\right)\right].\end{aligned} \tag{E.29}$$

Notice that with the choice of the learning rate

$$0 < \eta \le \min\left\{\frac{\tau}{24d_{\mathsf{max}}^2\|A\|_\infty^2(L+1)}, \frac{\zeta-1}{\zeta\tau}\right\},$$

we have

$$\frac{8(L+1)\eta d_{\mathsf{max}}^2\|A\|_\infty^2}{\tau} \le \frac{1}{2}$$

and

$$(1-\eta\tau)^{t+1}\mathsf{KL}\left(\pi_\tau^\star \,\|\, \pi^{(0)}\right) \ge \frac{1}{2}\sum_{l=0}^{t}(1-\eta\tau)^{t-l}\mathbb{E}\left[\Psi^{(l)}\right]$$

by combining (E.27) and (E.24). It follows that

$$\mathbb{E}\left[\Psi^{(t)}\right] \le 2(1-\eta\tau)^{t+1}\mathsf{KL}\left(\pi_\tau^\star \,\|\, \pi^{(0)}\right)$$

and

$$\begin{aligned}\mathbb{E}\left[\sum_{l=0}^{t-1} E(t-l)\Psi^{(l)}\right] &\overset{\text{(i)}}{\leq} \mathbb{E}\left[\sum_{l=0}^{t-1}(1-\eta\tau)^{t-l}\Psi^{(l)}\cdot E(t-l)\zeta^{t-l}\right]\\ &\leq \mathbb{E}\left[\sum_{l=0}^{t-1}(1-\eta\tau)^{t-l}\Psi^{(l)}\sum_{l=0}^{t-1} E(t-l)\zeta^{t-l}\right]\\ &\overset{\text{(ii)}}{\leq} 2L(1-\eta\tau)^{t+1}\mathsf{KL}\left(\pi_\tau^\star \,\|\, \pi^{(0)}\right),\end{aligned}$$

where (i) is by the bound $(1-\eta\tau)^{-1} \leq \zeta$ and (ii) uses the definition of L in Assumption 4. Plugging the above inequalities into (E.29) leads to

$$\begin{aligned}&(1-\eta\tau)\mathbb{E}\left[\mathsf{KL}\left(\pi_\tau^\star \,\|\, \pi^{(t)}\right)\right] + (1-\eta\tau)^{t+1}\mathsf{KL}\left(\pi_\tau^\star \,\|\, \pi^{(0)}\right)\\ &\geq \mathbb{E}\left[\mathsf{KL}\left(\pi_\tau^\star \,\|\, \overline{\pi}^{(t+1)}\right)\right] + \frac{\eta\tau}{2}\mathbb{E}\left[\mathsf{KL}\left(\overline{\pi}^{(t+1)} \,\|\, \pi_\tau^\star\right)\right].\end{aligned}$$

Then from (E.24) we have

$$\begin{aligned}\mathbb{E}\left[\mathsf{KL}\left(\pi_\tau^\star \,\|\, \overline{\pi}^{(t+1)}\right)\right] &\leq \mathbb{E}\left[\mathsf{KL}\left(\pi_\tau^\star \,\|\, \overline{\pi}^{(t+1)}\right)\right] + \frac{\eta\tau}{2}\mathbb{E}\left[\mathsf{KL}\left(\overline{\pi}^{(t+1)} \,\|\, \pi_\tau^\star\right)\right]\\ &\leq (1-\eta\tau)^{t+1}\mathsf{KL}\left(\pi_\tau^\star \,\|\, \pi^{(0)}\right) + (1-\eta\tau)^{t+1}\mathsf{KL}\left(\pi_\tau^\star \,\|\, \pi^{(0)}\right)\\ &= 2(1-\eta\tau)^{t+1}\mathsf{KL}\left(\pi_\tau^\star \,\|\, \pi^{(0)}\right). \qquad \text{(E.30)}\end{aligned}$$

Bounding the QRE-Gap. Combining (E.11) and (E.30), we have

$$\begin{aligned}\mathbb{E}\left[\texttt{QRE-Gap}_\tau(\overline{\pi}^{(t+1)})\right] &\leq \tau\mathbb{E}\left[\mathsf{KL}\left(\overline{\pi}^{(t+1)} \,\|\, \pi_\tau^\star\right)\right] + \frac{d_{\mathsf{max}}^2\|A\|_\infty^2}{\tau}\mathbb{E}\left[\mathsf{KL}\left(\pi_\tau^\star \,\|\, \overline{\pi}^{(t+1)}\right)\right]\\ &\leq \frac{2}{\eta}\left(\frac{\eta\tau}{2}\mathbb{E}\left[\mathsf{KL}\left(\overline{\pi}^{(t+1)} \,\|\, \pi_\tau^\star\right)\right] + \mathbb{E}\left[\mathsf{KL}\left(\pi_\tau^\star \,\|\, \overline{\pi}^{(t+1)}\right)\right]\right)\\ &\leq \frac{4(1-\eta\tau)^{t+1}}{\eta}\mathsf{KL}\left(\pi_\tau^\star \,\|\, \pi^{(0)}\right),\end{aligned}$$

where the second line uses the learning rate bound

$$\frac{2}{\eta} > \frac{24d_{\mathsf{max}}^2\|A\|_\infty^2(L+1)}{\tau} > \frac{d_{\mathsf{max}}^2\|A\|_\infty^2}{\tau}.$$

E.2 Proof for two-timescale OMWU (Section 6.3)

E.2.1 Proof of Theorem 10

Bounding $\mathsf{KL}\left(\pi_\tau^\star \,\|\, \pi^{(t)}\right)$. For notational convenience, we set $\pi^{(t)} = \overline{\pi}^{(t)} = \pi^{(0)}$ for $t < 0$. The following lemma parallels Lemma 36 by focusing on delayed feedbacks. The proof is postponed to Appendix E.3.4.

Lemma 38. *Assuming constant delays $\gamma_i^{(t)} = \gamma$, the iterates of OMWU based on the update rule (6.13) satisfy*

$$\left\langle \log\pi^{(t+1)} - (1-\eta\tau)\log\pi^{(t)} - \eta\tau\log\pi_\tau^\star,\, \overline{\pi}^{(t-\gamma+1)} - \pi_\tau^\star\right\rangle = 0.$$

By following a similar argument in (E.3), we conclude that

$$\begin{aligned}\mathsf{KL}\left(\pi_\tau^\star \,\|\, \pi^{(t+1)}\right) &= (1-\eta\tau)\mathsf{KL}\left(\pi_\tau^\star \,\|\, \pi^{(t)}\right) - (1-\eta\tau)\mathsf{KL}\left(\overline{\pi}^{(t-\gamma+1)} \,\|\, \pi^{(t)}\right) - \mathsf{KL}\left(\pi^{(t+1)} \,\|\, \overline{\pi}^{(t-\gamma+1)}\right) \\ &\quad + \left\langle \log \overline{\pi}^{(t-\gamma+1)} - \log \pi^{(t+1)},\, \overline{\pi}^{(t-\gamma+1)} - \pi^{(t+1)} \right\rangle - \eta\tau\mathsf{KL}\left(\overline{\pi}^{(t-\gamma+1)} \,\|\, \pi_\tau^\star\right).\end{aligned} \tag{E.31}$$

It boils down to control the term $-\left\langle \log \overline{\pi}^{(t-\gamma+1)} - \log \pi^{(t+1)},\, \overline{\pi}^{(t-\gamma+1)} - \pi^{(t+1)} \right\rangle$. When $t \geq \gamma$, by taking logarithm on the both sides of the update rules (6.9) and (6.13), we have

$$\log \overline{\pi}_i^{(t-\gamma+1)} \overset{\mathbf{1}}{=} (1-\overline{\eta}\tau)\log \pi_i^{(t-\gamma)} + \overline{\eta} A_i \overline{\pi}^{(t-2\gamma)}$$

and

$$\begin{aligned}\log \pi_i^{(t+1)} &\overset{\mathbf{1}}{=} (1-\eta\tau)\log \pi_i^{(t)} + \eta A_i \overline{\pi}^{(t-\gamma+1)} \\ &\overset{\mathbf{1}}{=} (1-\eta\tau)^{\gamma+1}\log \pi_i^{(t-\gamma)} + \eta\sum_{l=0}^{\gamma}(1-\eta\tau)^l A_i \overline{\pi}^{(t-\gamma-l+1)}.\end{aligned}$$

Subtracting the above equalities and taking inner product with $\overline{\pi}_i^{(t-\gamma+1)} - \pi_i^{(t+1)}$ gives

$$\begin{aligned}&\left\langle \log \overline{\pi}_i^{(t-\gamma+1)} - \log \pi_i^{(t+1)}, \overline{\pi}_i^{(t-\gamma+1)} - \pi_i^{(t+1)} \right\rangle \\ &= \eta\sum_{l=0}^{\gamma}(1-\eta\tau)^l \left\langle \overline{\pi}_i^{(t-\gamma+1)} - \pi_i^{(t+1)}, A_i(\overline{\pi}^{(t-2\gamma)} - \overline{\pi}^{(t-\gamma-l+1)}) \right\rangle,\end{aligned}$$

where the $\log \pi_i^{(t-\gamma)}$ terms cancel out due to the choice $1-\overline{\eta}\tau = (1-\eta\tau)^{\gamma+1}$. Summing over $i \in V$,

$$\begin{aligned}&\left\langle \log \overline{\pi}^{(t-\gamma+1)} - \log \pi^{(t+1)},\, \overline{\pi}^{(t-\gamma+1)} - \pi^{(t+1)} \right\rangle \\ &= \eta\sum_{i\in V}\sum_{l=0}^{\gamma}(1-\eta\tau)^l \left\langle \overline{\pi}_i^{(t-\gamma+1)} - \pi_i^{(t+1)},\, A_i(\overline{\pi}^{(t-2\gamma)} - \overline{\pi}^{(t-\gamma-l+1)}) \right\rangle \\ &\leq \eta\,\|A\|_\infty \sum_{(i,j)\in E}\sum_{l=0}^{\gamma}(1-\eta\tau)^l \left\|\overline{\pi}_i^{(t-\gamma+1)} - \pi_i^{(t+1)}\right\|_1 \left\|\overline{\pi}_j^{(t-2\gamma)} - \overline{\pi}_j^{(t-\gamma-l+1)}\right\|_1.\end{aligned} \tag{E.32}$$

Using the triangle inequality, we can bound $\left\|\overline{\pi}^{(t-2\gamma)} - \overline{\pi}^{(t-\gamma-l+1)}\right\|_1$ as

$$\begin{aligned}\left\|\overline{\pi}^{(t-2\gamma)} - \overline{\pi}^{(t-\gamma-l+1)}\right\|_1 &\leq \sum_{l_1=t-\gamma}^{t-l}\left\|\overline{\pi}_i^{(l_1-\gamma)} - \overline{\pi}_j^{(l_1-\gamma+1)}\right\|_1 \\ &\leq \sum_{l_1=t-\gamma}^{t-l}\left(\left\|\overline{\pi}_i^{(l_1-\gamma)} - \pi_i^{(l_1)}\right\|_1 + \left\|\overline{\pi}_j^{(l_1-\gamma+1)} - \pi_j^{(l_1)}\right\|_1\right).\end{aligned}$$

Substitution of the bound into (E.32) yields

$$
\begin{aligned}
&\left\langle \log \overline{\pi}^{(t-\gamma+1)} - \log \pi^{(t+1)},\, \overline{\pi}^{(t-\gamma+1)} - \pi^{(t+1)} \right\rangle \\
&\le \eta \left\|A\right\|_\infty \sum_{(i,j)\in E} \sum_{l=0}^{\gamma} (1-\eta\tau)^l \sum_{l_1=t-\gamma}^{t-l} \left\|\overline{\pi}_i^{(t-\gamma+1)} - \pi_i^{(t+1)}\right\|_1 \left(\left\|\overline{\pi}_j^{(l_1-\gamma)} - \pi_j^{(l_1)}\right\|_1 + \left\|\overline{\pi}_j^{(l_1-\gamma+1)} - \pi_j^{(l_1)}\right\|_1 \right) \\
&= \eta \left\|A\right\|_\infty \sum_{(i,j)\in E} \sum_{l_1=t-\gamma}^{t} \sum_{l=0}^{t-l_1} (1-\eta\tau)^l \left\|\overline{\pi}_i^{(t-\gamma+1)} - \pi_i^{(t+1)}\right\|_1 \left(\left\|\overline{\pi}_j^{(l_1-\gamma)} - \pi_j^{(l_1)}\right\|_1 + \left\|\overline{\pi}_j^{(l_1-\gamma+1)} - \pi_j^{(l_1)}\right\|_1 \right) \\
&\le \frac{1}{2}\eta \left\|A\right\|_\infty \sum_{(i,j)\in E} \Bigg[2 \sum_{l_1=t-\gamma}^{t} \sum_{l=0}^{t-l_1} (1-\eta\tau)^l \left\|\overline{\pi}_i^{(t-\gamma+1)} - \pi_i^{(t+1)}\right\|_1^2 \\
&\qquad + \sum_{l_1=t-\gamma}^{t} \sum_{l=0}^{t-l_1} (1-\eta\tau)^l \left(\left\|\overline{\pi}_j^{(l_1-\gamma)} - \pi_j^{(l_1)}\right\|_1^2 + \left\|\overline{\pi}_j^{(l_1-\gamma+1)} - \pi_j^{(l_1)}\right\|_1^2 \right) \Bigg] \\
&\le \eta d_{\mathsf{max}} \left\|A\right\|_\infty \Bigg[2(\gamma+1)^2 \mathsf{KL}\left(\pi^{(t+1)} \,\|\, \overline{\pi}^{(t-\gamma+1)}\right) \\
&\qquad + \sum_{l_1=t-\gamma}^{t} \sum_{l=0}^{t-l_1} (1-\eta\tau)^l \left(\mathsf{KL}\left(\pi^{(l_1)} \,\|\, \overline{\pi}^{(l_1-\gamma)}\right) + \mathsf{KL}\left(\overline{\pi}^{(l_1-\gamma+1)} \,\|\, \pi^{(l_1)}\right) \right) \Bigg].
\end{aligned} \tag{E.33}
$$

Plugging the above inequality into (E.31) and recursively applying the inequality gives

$$
\begin{aligned}
&\mathsf{KL}\left(\pi_\tau^\star \,\|\, \pi^{(t+1)}\right) + \mathsf{KL}\left(\pi^{(t+1)} \,\|\, \overline{\pi}^{(t-\gamma+1)}\right) + \eta\tau \mathsf{KL}\left(\overline{\pi}^{(t-\gamma+1)} \,\|\, \pi_\tau^\star\right) \\
&\le (1-\eta\tau)^{t+1-\gamma} \mathsf{KL}\left(\pi_\tau^\star \,\|\, \pi^{(\gamma)}\right) - \sum_{l_1=\gamma}^{t} (1-\eta\tau)^{t-l_1} \left(\mathsf{KL}\left(\pi^{(l_1+1)} \,\|\, \overline{\pi}^{(l_1-\gamma+1)}\right) + (1-\eta\tau)\mathsf{KL}\left(\overline{\pi}^{(l_1-\gamma+1)} \,\|\, \pi^{(l_1)}\right) \right) \\
&\quad + \eta d_{\mathsf{max}} \left\|A\right\|_\infty \Bigg[2(\gamma+1)^2 \sum_{l_1=\gamma}^{t} (1-\eta\tau)^{t-l_1} \mathsf{KL}\left(\pi^{(l_1+1)} \,\|\, \overline{\pi}^{(l_1-\gamma+1)}\right) \\
&\quad + \sum_{t_2=\gamma}^{t} (1-\eta\tau)^{t-l_2} \sum_{l_1=l_2-\gamma}^{l_2} \sum_{l=0}^{l_2-l_1} (1-\eta\tau)^l \left(\mathsf{KL}\left(\pi^{(l_1)} \,\|\, \overline{\pi}^{(l_1-\gamma)}\right) + \mathsf{KL}\left(\overline{\pi}^{(l_1-\gamma+1)} \,\|\, \pi^{(l_1)}\right) \right) \Bigg] \\
&\overset{\text{(i)}}{\le} (1-\eta\tau)^{t+1-\gamma} \mathsf{KL}\left(\pi_\tau^\star \,\|\, \pi^{(\gamma)}\right) - \sum_{l_1=\gamma}^{t} (1-\eta\tau)^{t-l_1} \left(\mathsf{KL}\left(\pi^{(l_1+1)} \,\|\, \overline{\pi}^{(l_1-\gamma+1)}\right) + (1-\eta\tau)\mathsf{KL}\left(\overline{\pi}^{(l_1-\gamma+1)} \,\|\, \pi^{(l_1)}\right) \right) \\
&\quad + 2(\gamma+1)^2 \eta d_{\mathsf{max}} \left\|A\right\|_\infty \sum_{l_1=\gamma}^{t} (1-\eta\tau)^{t-l_1} \mathsf{KL}\left(\pi^{(l_1+1)} \,\|\, \overline{\pi}^{(l_1-\gamma+1)}\right) \\
&\quad + 2(\gamma+1)^2 \eta d_{\mathsf{max}} \left\|A\right\|_\infty \sum_{l_1=0}^{t} (1-\eta\tau)^{t-l_1} \left(\mathsf{KL}\left(\pi^{(l_1+1)} \,\|\, \overline{\pi}^{(l_1-\gamma+1)}\right) + (1-\eta\tau)\mathsf{KL}\left(\overline{\pi}^{(l_1-\gamma+1)} \,\|\, \pi^{(l_1)}\right) \right) \\
&\overset{\text{(ii)}}{\le} (1-\eta\tau)^{t+1-\gamma} \mathsf{KL}\left(\pi_\tau^\star \,\|\, \pi^{(\gamma)}\right) \\
&\quad + 2(\gamma+1)^2 \eta d_{\mathsf{max}} \left\|A\right\|_\infty \sum_{l_1=0}^{\gamma-1} (1-\eta\tau)^{t-l_1} \left(\mathsf{KL}\left(\pi^{(l_1+1)} \,\|\, \overline{\pi}^{(l_1-\gamma+1)}\right) + (1-\eta\tau)\mathsf{KL}\left(\overline{\pi}^{(l_1-\gamma+1)} \,\|\, \pi^{(l_1)}\right) \right),
\end{aligned} \tag{E.34}
$$

where (i) results from basic calculation

$$
\begin{aligned}
&\sum_{t_2=\gamma}^{t}(1-\eta\tau)^{t-l_2}\sum_{l_1=l_2-\gamma}^{l_2}\sum_{l=0}^{l_2-l_1}(1-\eta\tau)^{l}\Big(\mathsf{KL}\left(\pi^{(l_1)}\,\|\,\overline{\pi}^{(l_1-\gamma)}\right)+\mathsf{KL}\left(\overline{\pi}^{(l_1-\gamma+1)}\,\|\,\pi^{(l_1)}\right)\Big)\\
&=\sum_{l_1=0}^{t}(1-\eta\tau)^{t-l_1}\sum_{l_2=l_1}^{l_1+\gamma}\sum_{l=0}^{l_2-l_1}(1-\eta\tau)^{l_1-l_2+l}\Big(\mathsf{KL}\left(\pi^{(l_1)}\,\|\,\overline{\pi}^{(l_1-\gamma)}\right)+\mathsf{KL}\left(\overline{\pi}^{(l_1-\gamma+1)}\,\|\,\pi^{(l_1)}\right)\Big)\\
&=\sum_{l_1=0}^{t}(1-\eta\tau)^{t-l_1}\sum_{l'=0}^{\gamma}\sum_{l=0}^{l'}(1-\eta\tau)^{l-l'}\Big(\mathsf{KL}\left(\pi^{(l_1)}\,\|\,\overline{\pi}^{(l_1-\gamma)}\right)+\mathsf{KL}\left(\overline{\pi}^{(l_1-\gamma+1)}\,\|\,\pi^{(l_1)}\right)\Big)\\
&\le\sum_{l_1=0}^{t}(1-\eta\tau)^{t-l_1}(\gamma+1)^2\Big(1-\frac{1}{2(\gamma+1)}\Big)^{-(\gamma+1)}(1-\eta\tau)\Big(\mathsf{KL}\left(\pi^{(l_1)}\,\|\,\overline{\pi}^{(l_1-\gamma)}\right)+\mathsf{KL}\left(\overline{\pi}^{(l_1-\gamma+1)}\,\|\,\pi^{(l_1)}\right)\Big)\\
&\le 2(\gamma+1)^2\sum_{l_1=0}^{t}(1-\eta\tau)^{t-l_1}(1-\eta\tau)\Big(\mathsf{KL}\left(\pi^{(l_1)}\,\|\,\overline{\pi}^{(l_1-\gamma)}\right)+\mathsf{KL}\left(\overline{\pi}^{(l_1-\gamma+1)}\,\|\,\pi^{(l_1)}\right)\Big)\\
&\le 2(\gamma+1)^2\sum_{l_1=0}^{t}(1-\eta\tau)^{t-l_1}\Big(\mathsf{KL}\left(\pi^{(l_1+1)}\,\|\,\overline{\pi}^{(l_1-\gamma+1)}\right)+(1-\eta\tau)\mathsf{KL}\left(\overline{\pi}^{(l_1-\gamma+1)}\,\|\,\pi^{(l_1)}\right)\Big)
\end{aligned}
$$

and (ii) is due to $\eta\le\min\left\{\frac{1}{2\tau(\gamma+1)},\frac{1}{5d_{\mathsf{max}}\|A\|_\infty(\gamma+1)^2}\right\}$. To proceed, we introduce the following lemma concerning the error $\mathsf{KL}\left(\pi_\tau^\star\,\|\,\pi^{(\gamma)}\right)$, with the proof postponed to Appendix E.3.5.

Lemma 39. *With constant delays $\gamma_i^{(t)}=\gamma$, the iterates of OMWU based on the update rule (6.13) satisfy*

$$
\begin{aligned}
&\mathsf{KL}\left(\pi_\tau^\star\,\|\,\pi^{(\gamma)}\right)\\
&\le(1-\eta\tau)^{\gamma}\mathsf{KL}\left(\pi_\tau^\star\,\|\,\pi^{(0)}\right)-\sum_{l_1=0}^{\gamma-1}(1-\eta\tau)^{\gamma-1-l_1}\Big(\mathsf{KL}\left(\pi^{(l_1+1)}\,\|\,\overline{\pi}^{(l_1-\gamma+1)}\right)+(1-\eta\tau)\mathsf{KL}\left(\overline{\pi}^{(l_1-\gamma+1)}\,\|\,\pi^{(l_1)}\right)\Big)\\
&\quad+2\eta\gamma^2 d_{\mathsf{max}}\|A\|_\infty.
\end{aligned}
$$

With the lemma above in mind, we can continue to bound (E.34) by

$$
\begin{aligned}
&\mathsf{KL}\left(\pi_\tau^\star\,\|\,\pi^{(t+1)}\right)\\
&\le(1-\eta\tau)^{t+1}\mathsf{KL}\left(\pi_\tau^\star\,\|\,\pi^{(0)}\right)+2(1-\eta\tau)^{t+1-\gamma}\eta\gamma^2 d_{\mathsf{max}}\|A\|_\infty\\
&\quad-\sum_{l_1=0}^{\gamma-1}(1-\eta\tau)^{t-l_1}\Big(\mathsf{KL}\left(\pi^{(l_1+1)}\,\|\,\overline{\pi}^{(l_1-\gamma+1)}\right)+(1-\eta\tau)\mathsf{KL}\left(\overline{\pi}^{(l_1-\gamma+1)}\,\|\,\pi^{(l_1)}\right)\Big)\\
&\quad+2(\gamma+1)^2\eta d_{\mathsf{max}}\|A\|_\infty\sum_{l_1=0}^{\gamma-1}(1-\eta\tau)^{t-l_1}\Big(\mathsf{KL}\left(\pi^{(l_1+1)}\,\|\,\overline{\pi}^{(l_1-\gamma+1)}\right)+(1-\eta\tau)\mathsf{KL}\left(\overline{\pi}^{(l_1-\gamma+1)}\,\|\,\pi^{(l_1)}\right)\Big)\\
&\le(1-\eta\tau)^{t+1}\mathsf{KL}\left(\pi_\tau^\star\,\|\,\pi^{(0)}\right)+(1-\eta\tau)^{t+1-\gamma}.
\end{aligned}
$$

Bounding $\mathsf{KL}\left(\pi_\tau^\star \,\|\, \overline{\pi}^{(t-\gamma+1)}\right)$. By definition of KL divergence, we have

$$\begin{aligned}
&\mathsf{KL}\left(\pi_\tau^\star \,\|\, \overline{\pi}^{(t-\gamma+1)}\right) \\
&= \mathsf{KL}\left(\pi_\tau^\star \,\|\, \pi^{(t+1)}\right) + \left\langle \pi_\tau^\star, \log \pi^{(t+1)} - \log \overline{\pi}^{(t-\gamma+1)} \right\rangle \\
&= \mathsf{KL}\left(\pi_\tau^\star \,\|\, \pi^{(t+1)}\right) + \mathsf{KL}\left(\pi^{(t+1)} \,\|\, \overline{\pi}^{(t-\gamma+1)}\right) + \left\langle \pi_\tau^\star - \pi^{(t+1)}, \log \pi^{(t+1)} - \log \overline{\pi}^{(t-\gamma+1)} \right\rangle. \qquad \text{(E.35)}
\end{aligned}$$

It remains to control the term $\left\langle \pi_\tau^\star - \pi^{(t+1)}, \log \pi^{(t+1)} - \log \overline{\pi}^{(t-\gamma+1)} \right\rangle$. By following a similar argument in (E.33), we have

$$\begin{aligned}
&\left\langle \pi_\tau^\star - \pi^{(t+1)}, \log \pi^{(t+1)} - \log \overline{\pi}^{(t-\gamma+1)} \right\rangle \\
&= \eta \sum_{i \in V} \sum_{l=0}^{\gamma} (1-\eta\tau)^l \left\langle \pi_{i,\tau}^\star - \pi_i^{(t+1)},\, A_i(\overline{\pi}^{(t-2\gamma)} - \overline{\pi}^{(t-\gamma-l+1)}) \right\rangle \\
&\le \eta \left\|A\right\|_\infty \sum_{(i,j) \in E} \sum_{l=0}^{\gamma} (1-\eta\tau)^l \left\|\pi_{i,\tau}^\star - \pi_i^{(t+1)}\right\|_1 \left\|\overline{\pi}_j^{(t-2\gamma)} - \overline{\pi}_j^{(t-\gamma-l+1)}\right\|_1 \\
&\le \eta \left\|A\right\|_\infty \sum_{(i,j) \in E} \sum_{l=0}^{\gamma} (1-\eta\tau)^l \sum_{l_1=t-\gamma}^{t-l} \left\|\pi_{i,\tau}^\star - \pi_i^{(t+1)}\right\|_1 \left(\left\|\overline{\pi}_j^{(l_1-\gamma)} - \pi_j^{(l_1)}\right\|_1 + \left\|\overline{\pi}_j^{(l_1-\gamma+1)} - \pi_j^{(l_1)}\right\|_1 \right) \\
&= \eta \left\|A\right\|_\infty \sum_{(i,j) \in E} \sum_{l_1=t-\gamma}^{t} \sum_{l=0}^{t-l_1} (1-\eta\tau)^l \left\|\pi_{i,\tau}^\star - \pi_i^{(t+1)}\right\|_1 \left(\left\|\overline{\pi}_j^{(l_1-\gamma)} - \pi_j^{(l_1)}\right\|_1 + \left\|\overline{\pi}_j^{(l_1-\gamma+1)} - \pi_j^{(l_1)}\right\|_1 \right) \\
&\le \eta d_{\mathsf{max}} \left\|A\right\|_\infty \Bigg[2(\gamma+1)^2 \mathsf{KL}\left(\pi_\tau^\star \,\|\, \pi^{(t+1)}\right) \\
&\qquad + \sum_{l_1=t-\gamma}^{t} \sum_{l=0}^{t-l_1} (1-\eta\tau)^l \left(\mathsf{KL}\left(\pi^{(l_1)} \,\|\, \overline{\pi}^{(l_1-\gamma)}\right) + \mathsf{KL}\left(\overline{\pi}^{(l_1-\gamma+1)} \,\|\, \pi^{(l_1)}\right) \right) \Bigg].
\end{aligned}$$

Substitution of the above inequality into (E.35) yields

$$
\begin{aligned}
&\mathsf{KL}\left(\pi_\tau^\star \,\|\, \overline{\pi}^{(t-\gamma+1)}\right) + \eta\tau \mathsf{KL}\left(\overline{\pi}^{(t-\gamma+1)} \,\|\, \pi_\tau^\star\right) \\
&= (1+2(\gamma+1)^2\eta d_{\mathsf{max}})\mathsf{KL}\left(\pi_\tau^\star \,\|\, \pi^{(t+1)}\right) + \mathsf{KL}\left(\pi^{(t+1)} \,\|\, \overline{\pi}^{(t-\gamma+1)}\right) + \eta\tau \mathsf{KL}\left(\overline{\pi}^{(t-\gamma+1)} \,\|\, \pi_\tau^\star\right) \\
&\qquad + \eta d_{\mathsf{max}} \|A\|_\infty \sum_{l_1=t-\gamma}^{t} \sum_{l=0}^{t-l_1} (1-\eta\tau)^l \left(\mathsf{KL}\left(\pi^{(l_1)} \,\|\, \overline{\pi}^{(l_1-\gamma)}\right) + \mathsf{KL}\left(\overline{\pi}^{(l_1-\gamma+1)} \,\|\, \pi^{(l_1)}\right)\right) \\
&\overset{\text{(i)}}{\leq} 2\Big(\mathsf{KL}\left(\pi_\tau^\star \,\|\, \pi^{(t+1)}\right) + \mathsf{KL}\left(\pi^{(t+1)} \,\|\, \overline{\pi}^{(t-\gamma+1)}\right) + \eta\tau \mathsf{KL}\left(\overline{\pi}^{(t-\gamma+1)} \,\|\, \pi_\tau^\star\right)\Big) \\
&\qquad + 2(\gamma+1)\eta d_{\mathsf{max}} \|A\|_\infty \sum_{l_1=0}^{t} (1-\eta\tau)^{t-l_1} \left(\mathsf{KL}\left(\pi^{(l_1+1)} \,\|\, \overline{\pi}^{(l_1-\gamma+1)}\right) + (1-\eta\tau)\mathsf{KL}\left(\overline{\pi}^{(l_1-\gamma+1)} \,\|\, \pi^{(l_1)}\right)\right) \\
&\overset{\text{(ii)}}{\leq} 2(1-\eta\tau)^{t+1-\gamma}\mathsf{KL}\left(\pi_\tau^\star \,\|\, \pi^{(\gamma)}\right) - 2\sum_{l_1=\gamma}^{t} (1-\eta\tau)^{t-l_1} \left(\mathsf{KL}\left(\pi^{(l_1+1)} \,\|\, \overline{\pi}^{(l_1-\gamma+1)}\right) + (1-\eta\tau)\mathsf{KL}\left(\overline{\pi}^{(l_1-\gamma+1)} \,\|\, \pi^{(l_1)}\right)\right) \\
&\quad + 4(\gamma+1)^2\eta d_{\mathsf{max}} \|A\|_\infty \sum_{l_1=\gamma}^{t} (1-\eta\tau)^{t-l_1} \mathsf{KL}\left(\pi^{(l_1+1)} \,\|\, \overline{\pi}^{(l_1-\gamma+1)}\right) \\
&\quad + 6(\gamma+1)^2\eta d_{\mathsf{max}} \|A\|_\infty \sum_{l_1=0}^{t} (1-\eta\tau)^{t-l_1} \Big(\mathsf{KL}\left(\pi^{(l_1+1)} \,\|\, \overline{\pi}^{(l_1-\gamma+1)}\right) + (1-\eta\tau)\mathsf{KL}\left(\overline{\pi}^{(l_1-\gamma+1)} \,\|\, \pi^{(l_1)}\right)\Big) \\
&\leq 2(1-\eta\tau)^{t+1-\gamma}\mathsf{KL}\left(\pi_\tau^\star \,\|\, \pi^{(\gamma)}\right) \\
&\quad + 6(\gamma+1)^2\eta d_{\mathsf{max}} \|A\|_\infty \sum_{l_1=0}^{\gamma-1} (1-\eta\tau)^{t-l_1} \Big(\mathsf{KL}\left(\pi^{(l_1+1)} \,\|\, \overline{\pi}^{(l_1-\gamma+1)}\right) + (1-\eta\tau)\mathsf{KL}\left(\overline{\pi}^{(l_1-\gamma+1)} \,\|\, \pi^{(l_1)}\right)\Big),
\end{aligned}
$$

where (i) results from

$$
\begin{aligned}
&\sum_{l_1=t-\gamma}^{t} \sum_{l=0}^{t-l_1} (1-\eta\tau)^l \left(\mathsf{KL}\left(\pi^{(l_1)} \,\|\, \overline{\pi}^{(l_1-\gamma)}\right) + \mathsf{KL}\left(\overline{\pi}^{(l_1-\gamma+1)} \,\|\, \pi^{(l_1)}\right)\right) \\
&= \sum_{l_1=t-\gamma}^{t} (1-\eta\tau)^{t-l_1} \sum_{l=0}^{t-l_1} (1-\eta\tau)^{l+l_1-t} \left(\mathsf{KL}\left(\pi^{(l_1)} \,\|\, \overline{\pi}^{(l_1-\gamma)}\right) + \mathsf{KL}\left(\overline{\pi}^{(l_1-\gamma+1)} \,\|\, \pi^{(l_1)}\right)\right) \\
&\leq \sum_{l_1=t-\gamma}^{t} (1-\eta\tau)^{t-l_1} (\gamma+1)(1-\eta\tau)^{-(\gamma+1)}(1-\eta\tau) \left(\mathsf{KL}\left(\pi^{(l_1)} \,\|\, \overline{\pi}^{(l_1-\gamma)}\right) + \mathsf{KL}\left(\overline{\pi}^{(l_1-\gamma+1)} \,\|\, \pi^{(l_1)}\right)\right) \\
&\leq 2(\gamma+1) \sum_{l_1=0}^{t} (1-\eta\tau)^{t-l_1} \left(\mathsf{KL}\left(\pi^{(l_1+1)} \,\|\, \overline{\pi}^{(l_1-\gamma+1)}\right) + (1-\eta\tau)\mathsf{KL}\left(\overline{\pi}^{(l_1-\gamma+1)} \,\|\, \pi^{(l_1)}\right)\right).
\end{aligned}
$$

and (ii) is due to the bound established in (E.34). Finally, applying Lemma 39 yields

$$
\begin{aligned}
&\mathsf{KL}\left(\pi_\tau^\star \,\|\, \overline{\pi}^{(t-\gamma+1)}\right) + \eta\tau \mathsf{KL}\left(\overline{\pi}^{(t-\gamma+1)} \,\|\, \pi_\tau^\star\right) \\
&\le 2(1-\eta\tau)^{t+1}\mathsf{KL}\left(\pi_\tau^\star \,\|\, \pi^{(0)}\right) + 4(1-\eta\tau)^{t+1-\gamma}\eta\gamma^2 d_{\mathsf{max}} \|A\|_\infty \\
&\quad - 2\sum_{l_1=0}^{\gamma-1}(1-\eta\tau)^{t-l_1}\left(\mathsf{KL}\left(\pi^{(l_1+1)} \,\|\, \overline{\pi}^{(l_1-\gamma+1)}\right) + (1-\eta\tau)\mathsf{KL}\left(\overline{\pi}^{(l_1-\gamma+1)} \,\|\, \pi^{(l_1)}\right)\right) \\
&\quad + 6(\gamma+1)^2\eta d_{\mathsf{max}} \|A\|_\infty \sum_{l_1=0}^{\gamma-1}(1-\eta\tau)^{t-l_1}\left(\mathsf{KL}\left(\pi^{(l_1+1)} \,\|\, \overline{\pi}^{(l_1-\gamma+1)}\right) + (1-\eta\tau)\mathsf{KL}\left(\overline{\pi}^{(l_1-\gamma+1)} \,\|\, \pi^{(l_1)}\right)\right) \\
&\le 2(1-\eta\tau)^{t+1}\mathsf{KL}\left(\pi_\tau^\star \,\|\, \pi^{(0)}\right) + 2(1-\eta\tau)^{t+1-\gamma}. \qquad \text{(E.36)}
\end{aligned}
$$

Bounding the QRE gap. With Lemma 37, we have

$$
\begin{aligned}
\texttt{QRE-Gap}_\tau(\overline{\pi}^{(t-\gamma+1)}) &\le \frac{d_{\mathsf{max}}^2 \|A\|_\infty^2}{\tau}\mathsf{KL}\left(\pi_\tau^\star \,\|\, \overline{\pi}^{(t-\gamma+1)}\right) + \tau\mathsf{KL}\left(\overline{\pi}^{(t-\gamma+1)} \,\|\, \pi_\tau^\star\right) \\
&\le \max\left\{\frac{d_{\mathsf{max}}^2 \|A\|_\infty^2}{\tau}, \frac{1}{\eta}\right\}\left(\mathsf{KL}\left(\pi_\tau^\star \,\|\, \overline{\pi}^{(t-\gamma+1)}\right) + \eta\tau\mathsf{KL}\left(\overline{\pi}^{(t-\gamma+1)} \,\|\, \pi_\tau^\star\right)\right) \\
&\le 2\max\left\{\frac{d_{\mathsf{max}}^2 \|A\|_\infty^2}{\tau}, \frac{1}{\eta}\right\}\left((1-\eta\tau)^{t+1}\mathsf{KL}\left(\pi_\tau^\star \,\|\, \pi^{(0)}\right) + (1-\eta\tau)^{t+1-\gamma}\right),
\end{aligned}
$$

where the last step results from (E.36).

E.2.2 Proof of Theorem 11

Bounding the term $\mathsf{KL}\left(\pi_\tau^\star \,\|\, \pi^{(t)}\right)$. Recall that the update rule of $\pi_i^{(t)}(k)$ is given by

$$
\pi_i^{(t)}(k) \propto \pi_i^{(t-1)}(k)^{1-\eta\tau}\exp(\eta[A_i\overline{\pi}^{(\kappa_i^{(t)})}]_k). \qquad \text{(E.37)}
$$

We introduce an auxiliary variable $\widetilde{\pi}_i^{(t)}$:

$$
\widetilde{\pi}_i^{(t)}(k) \propto \pi_i^{(t-1)}(k)^{1-\widetilde{\eta}_i^{(t)}\tau}\exp\left(\widetilde{\eta}_i^{(t)}[A_i\overline{\pi}^{(\kappa_i^{(t)})}]_k\right), \qquad \text{(E.38)}
$$

which can be viewed as a conceptual alternative update of $\pi_i^{(t)}$ with a different step size $\widetilde{\eta}_i^{(t)} > 0$ satisfying

$$
(1-\widetilde{\eta}_i^{(t)}\tau)(1-\eta\tau)^{t-\kappa_i^{(t)}} = 1-\overline{\eta}\tau
$$

or equivalently

$$
1-\widetilde{\eta}_i^{(t)}\tau = (1-\eta\tau)^{\gamma+1-t+\kappa_i^{(t)}}.
$$

It directly follows that $\widetilde{\eta}_i^{(t)} \ge \eta$. Since $\kappa_i^{(t)} \le t$, we have $1-\widetilde{\eta}_i^{(t)}\tau \ge 1-(\gamma+1-t+\kappa_i^{(t)})\eta\tau \ge 1-(\gamma+1)\eta\tau$, which implies $\widetilde{\eta}_i^{(t)} \le (\gamma+1)\eta$. For notational convenience, we set $\widetilde{\pi}_i^{(t)} = \pi^{(0)}$, $\widetilde{\eta}_i^{(t)} = \eta$ and $\kappa_i^{(t)} = 0$ when $t \le 0$. The following lemma establishes a one-step analysis, with the proof postponed to Appendix E.3.6.

Lemma 40. *When $t \geq 1$, it holds that*

$$\mathsf{KL}\left(\pi_{i,\tau}^{\star} \,\|\, \pi_i^{(t)}\right) + \eta\tau \mathsf{KL}\left(\overline{\pi}_i^{(\kappa_i^{(t)})} \,\|\, \pi_{i,\tau}^{\star}\right) = (1-\eta\tau)\mathsf{KL}\left(\pi_{i,\tau}^{\star} \,\|\, \pi_i^{(t-1)}\right) - \eta(\overline{\pi}_i^{(\kappa_i^{(t)})} - \pi_{i,\tau}^{\star})^\top A_i(\overline{\pi}^{(\kappa_i^{(t)})} - \pi_\tau^{\star}) - \psi_i^{(t)} + \frac{\eta}{\widetilde{\eta}_i^{(t)}}\left\langle \log \overline{\pi}_i^{(\kappa_i^{(t)})} - \log \widetilde{\pi}_i^{(t)},\, \overline{\pi}_i^{(\kappa_i^{(t)})} - \widetilde{\pi}_i^{(t)} \right\rangle, \tag{E.39}$$

where

$$\psi_i^{(t)} := \Big(1 - \frac{\eta}{\widetilde{\eta}_i^{(t)}}\Big)\mathsf{KL}\left(\pi_i^{(t)} \,\|\, \pi_i^{(t-1)}\right) + \frac{\eta}{\widetilde{\eta}_i^{(t)}}\left[(1-\widetilde{\eta}_i^{(t)}\tau)\mathsf{KL}\Big(\overline{\pi}_i^{(\kappa_i^{(t)})} \,\|\, \pi_i^{(t-1)}\Big) + \mathsf{KL}\Big(\widetilde{\pi}_i^{(t)} \,\|\, \overline{\pi}_i^{(\kappa_i^{(t)})}\Big) + \mathsf{KL}\Big(\pi_i^{(t)} \,\|\, \widetilde{\pi}_i^{(t)}\Big)\right].$$

We proceed to control the term $\left\langle \log \overline{\pi}_i^{(\kappa_i^{(t)})} - \log \widetilde{\pi}_i^{(t)},\, \overline{\pi}_i^{(\kappa_i^{(t)})} - \widetilde{\pi}_i^{(t)} \right\rangle$. By definition, we have

$$\begin{aligned}
\log \widetilde{\pi}_i^{(t)} &\overset{\mathbf{1}}{=} (1-\widetilde{\eta}_i^{(t)}\tau)\log \pi_i^{(t-1)} + \widetilde{\eta}_i^{(t)} A_i \overline{\pi}^{(\kappa_i^{(t)})} \\
&\overset{\mathbf{1}}{=} (1-\widetilde{\eta}_i^{(t)}\tau)(1-\eta\tau)^{t-\kappa_i^{(t)}} \log \pi^{(\kappa_i^{(t)}-1)} \\
&\quad + \widetilde{\eta}_i^{(t)}\Big(A_i\overline{\pi}^{(\kappa_i^{(t)})} + \sum_{l=\kappa_i^{(t)}}^{t-1}(1-\widetilde{\eta}_i^{(t)}\tau)(1-\eta\tau)^{t-1-l}A_i\overline{\pi}^{(\kappa_i^{(l)})}\Big)
\end{aligned}$$

and

$$\log \overline{\pi}_i^{(\kappa_i^{(t)})} \overset{\mathbf{1}}{=} (1-\overline{\eta}\tau)\log \pi^{(\kappa_i^{(t)}-1)} + \overline{\eta} A_i \overline{\pi}^{(\kappa_i^{(\kappa_i^{(t)}-1)})}$$

when $\kappa_i^{(t)} \geq 1$. Subtracting the two equations yields

$$\begin{aligned}
\log \overline{\pi}_i^{(\kappa_i^{(t)})} - \log \widetilde{\pi}_i^{(t)} &\overset{\mathbf{1}}{=} \widetilde{\eta}_i^{(t)}\Big(A_i(\overline{\pi}^{(\kappa_i^{(\kappa_i^{(t)}-1)})} - \overline{\pi}^{(\kappa_i^{(t)})}) \\
&\qquad + \sum_{l=\kappa_i^{(t)}}^{t-1}(1-\widetilde{\eta}_i^{(t)}\tau)(1-\eta\tau)^{t-1-l}A_i(\overline{\pi}^{(\kappa_i^{(\kappa_i^{(t)}-1)})} - \overline{\pi}^{(\kappa_i^{(l)})})\Big),
\end{aligned} \tag{E.40}$$

where the $\log \pi^{(\kappa_i^{(t)}-1)}$ terms cancel out due to $(1-\widetilde{\eta}_i^{(t)}\tau)(1-\eta\tau)^{t-\kappa_i^{(t)}} = 1-\overline{\eta}\tau$. It follows that

$$\begin{aligned}
&\left\langle \log \overline{\pi}_i^{(\kappa_i^{(t)})} - \log \widetilde{\pi}_i^{(t)},\, \overline{\pi}_i^{(\kappa_i^{(t)})} - \widetilde{\pi}_i^{(t)} \right\rangle \\
&= \widetilde{\eta}_i^{(t)}\Big(\left\langle \overline{\pi}_i^{(\kappa_i^{(t)})} - \widetilde{\pi}_i^{(t)},\, A_i(\overline{\pi}^{(\kappa_i^{(\kappa_i^{(t)}-1)})} - \overline{\pi}^{(\kappa_i^{(t)})}) \right\rangle \\
&\qquad + \sum_{l=\kappa_i^{(t)}}^{t-1}(1-\widetilde{\eta}_i^{(t)}\tau)(1-\eta\tau)^{t-1-l}\left\langle \overline{\pi}_i^{(\kappa_i^{(t)})} - \widetilde{\pi}_i^{(t)},\, A_i(\overline{\pi}^{(\kappa_i^{(\kappa_i^{(t-1)})})} - \overline{\pi}^{(\kappa_i^{(l)})}) \right\rangle\Big) \\
&\leq \widetilde{\eta}_i^{(t)} \|A\|_\infty \big\|\overline{\pi}_i^{(\kappa_i^{(t)})} - \widetilde{\pi}_i^{(t)}\big\|_1 \sum_{j\in\mathcal{N}_i}\sum_{l=\kappa_i^{(t)}}^{t} \big\|\overline{\pi}_j^{(\kappa_i^{(l)})} - \overline{\pi}_j^{(\kappa_i^{(\kappa_i^{(t-1)})})}\big\|_1.
\end{aligned} \tag{E.41}$$

The next lemma establishes an upper bound on the term $\sum_{l=\kappa_i^{(t)}}^{t} \big\|\overline{\pi}_j^{(\kappa_i^{(l)})} - \overline{\pi}_j^{(\kappa_i^{(\kappa_i^{(t-1)})})}\big\|_1$, with the proof postponed to Appendix E.3.7.

Lemma 41. *Let $\nu_j(t)$ denote the time index when agent j receives the payoff from the t-th iteration, i.e., $\kappa_j^{(\nu_j(t))} = t$. For $t = 0$, we set $\nu_j(0)$ to an arbitrary index that satisfies $\kappa_j^{(\nu_j(0))} = 0$. When $t \geq 2\gamma + 1$, it holds that*

$$\sum_{l=\kappa_i^{(t)}}^{t} \big\|\overline{\pi}_j^{(\kappa_i^{(l)})} - \overline{\pi}_j^{(\kappa_i^{(\kappa_i^{(t-1)})})}\big\|_1 \leq 4\sqrt{2}(\gamma+1) \sum_{l=t-2\gamma}^{t+\gamma} \sqrt{\psi_j^{(l)}} + 2\sqrt{2}(\gamma+1)^2 \sqrt{\psi_j^{(\nu_j(\kappa_i^{(\kappa_i^{(t-1)})}))}},$$

Plugging Lemma 41 into (E.41) gives

$$\begin{aligned}
&\Big\langle \log \overline{\pi}_i^{(\kappa_i^{(t)})} - \log \widetilde{\pi}_i^{(t)},\, \overline{\pi}_i^{(\kappa_i^{(t)})} - \widetilde{\pi}_i^{(t)} \Big\rangle \\
&\leq \widetilde{\eta}_i^{(t)} \|A\|_\infty \big\|\overline{\pi}_i^{(\kappa_i^{(k)})} - \widetilde{\pi}_i^{(t)}\big\|_1 \sum_{j\in\mathcal{N}_i} \Big[4\sqrt{2}(\gamma+1) \sum_{l=t-2\gamma}^{t+\gamma} \sqrt{\psi_j^{(l)}} + 2\sqrt{2}(\gamma+1)^2 \sqrt{\psi_j^{(\nu_j(\kappa_i^{(\kappa_i^{(t-1)})}))}}\Big] \\
&\overset{\text{(i)}}{\leq} \frac{1}{2}\widetilde{\eta}_i^{(t)} \|A\|_\infty \Big\{ 14 d_{\mathsf{max}} (\gamma+1)^{3/2} \big\|\overline{\pi}_i^{(\kappa_i^{(t)})} - \widetilde{\pi}_i^{(t)}\big\|_1^2 \\
&\qquad + \sum_{j\in\mathcal{N}_i} \Big[8(\gamma+1)^{3/2} \sum_{l=t-2\gamma}^{t+\gamma} \psi_j^{(l)} + 4(\gamma+1)^{5/2} \psi_j^{(\nu_j(\kappa_i^{(\kappa_i^{(t-1)})}))}\Big]\Big\} \\
&\overset{\text{(ii)}}{\leq} \widetilde{\eta}_i^{(t)} \|A\|_\infty \Big\{ 14 d_{\mathsf{max}} (\gamma+1)^{5/2} \psi_i^{(t)} + \sum_{j\in\mathcal{N}_i} \Big[4(\gamma+1)^{3/2} \sum_{l=t-2\gamma}^{t+\gamma} \psi_j^{(l)} + 2(\gamma+1)^{5/2} \psi_j^{(\nu_j(\kappa_i^{(\kappa_i^{(t-1)})}))}\Big]\Big\},
\end{aligned} \tag{E.42}$$

where (i) results from Young's inequality

$$\big\|\overline{\pi}_i^{(\kappa_i^{(t)})} - \widetilde{\pi}_i^{(t)}\big\|_1 \sqrt{\psi_j^{(l)}} \leq \frac{1}{2}\Big(\frac{1}{\sqrt{2}(\gamma+1)^{1/2}} \big\|\overline{\pi}_i^{(\kappa_i^{(t)})} - \widetilde{\pi}_i^{(t)}\big\|_1^2 + \sqrt{2}(\gamma+1)^{1/2} \psi_j^{(l)}\Big)$$

and (ii) follows from $\Big\|\overline{\pi}_i^{(\kappa_i^{(t)})} - \widetilde{\pi}_i^{(t)}\Big\|_1^2 \leq 2\mathsf{KL}\Big(\widetilde{\pi}_i^{(t)} \,\|\, \overline{\pi}_i^{(\kappa_i^{(t)})}\Big) \leq 2(\gamma+1)\psi_i^{(t)}$. Plugging (E.42) into (E.39) and summing over $i \in V$ yields

$$\begin{aligned}
&\mathsf{KL}\left(\pi_\tau^\star \,\|\, \pi^{(t)}\right) + \eta\tau \sum_{i\in V} \mathsf{KL}\Big(\overline{\pi}_i^{(\kappa_i^{(t)})} \,\|\, \pi_\tau^\star\Big) \\
&\leq (1-\eta\tau)\mathsf{KL}\left(\pi_\tau^\star \,\|\, \pi^{(t-1)}\right) - \eta \sum_{i\in V} (\overline{\pi}_i^{(\kappa_i^{(t)})} - \pi_{i,\tau}^\star)^\top A_i (\overline{\pi}^{(\kappa_i^{(t)})} - \pi_\tau^\star) \\
&\qquad - (1 - 14\eta d_{\mathsf{max}} \|A\|_\infty (\gamma+1)^{5/2}) \sum_{i\in V} \psi_i^{(t)} + 2\eta \|A\|_\infty (\gamma+1)^{5/2} \sum_{(i,j)\in E} \psi_j^{(\nu_j(\kappa_i^{(\kappa_i^{(t-1)})}))} \\
&\qquad + 4\eta d_{\mathsf{max}} \|A\|_\infty (\gamma+1)^{3/2} \sum_{l=t-2\gamma}^{t+\gamma} \psi^{(l)},
\end{aligned} \tag{E.43}$$

where we denote $\sum_{i\in V}\psi_i^{(l)}$ by $\psi^{(l)}$ for notation simplicity. We then seek to sum the above equation over $t=2\gamma+1,\cdots,T$. Before proceeding, we note that

$$\sum_{t=2\gamma+1}^{T}\sum_{l=t-2\gamma}^{t+\gamma}\psi^{(l)}\leq\sum_{l=1}^{T+\gamma}\sum_{t=l-\gamma}^{l+2\gamma}\psi^{(l)}\leq 3(\gamma+1)\sum_{l=1}^{T+\gamma}\psi^{(l)},$$

and that

$$\sum_{t=2\gamma+1}^{T}\sum_{(i,j)\in E}\psi_j^{(\nu_j(\kappa_i^{(\kappa_i^{(t-1)})}))}\leq\sum_{(i,j)\in E}\sum_{t=0}^{T+\gamma-1}\psi_j^{(t)}\leq d_{\mathsf{max}}\sum_{t=1}^{T+\gamma-1}\psi^{(t)},$$

where the first step is due to the mapping $t\mapsto\nu_j(\kappa_i^{(\kappa_i^{(t-1)})})$ being injective when $t\geq 2\gamma+1$ (cf. Assumptions 5, 6). Note that $\psi_j^{(t)}=0$ when $t\leq 0$ and hence can be safely discarded. Taken together, we arrive at

$$\begin{aligned}
&\eta\tau\sum_{t=2\gamma+1}^{T}\mathsf{KL}\left(\pi_\tau^\star\,\|\,\pi^{(t)}\right)+\eta\tau\sum_{t=2\gamma+1}^{T}\sum_{i\in V}\mathsf{KL}\left(\overline{\pi}_i^{(\kappa_i^{(t)})}\,\|\,\pi_{i,\tau}^\star\right)\\
&\leq(1-\eta\tau)\mathsf{KL}\left(\pi_\tau^\star\,\|\,\pi^{(2\gamma)}\right)-\eta\sum_{t=2\gamma+1}^{T}\sum_{i\in V}(\overline{\pi}_i^{(\kappa_i^{(t)})}-\pi_{i,\tau}^\star)^\top A_i(\overline{\pi}^{(\kappa_i^{(t)})}-\pi_\tau^\star)\\
&\quad-\left(1-14\eta d_{\mathsf{max}}\left\|A\right\|_\infty(\gamma+1)^{5/2}\right)\sum_{t=2\gamma+1}^{T}\psi^{(t)}+12\eta d_{\mathsf{max}}\left\|A\right\|_\infty(\gamma+1)^{5/2}\sum_{l=1}^{T+\gamma}\psi^{(l)}\\
&\quad+2\eta d_{\mathsf{max}}\left\|A\right\|_\infty(\gamma+1)^{5/2}\sum_{t=1}^{T+\gamma-1}\psi^{(t)}\\
&\leq(1-\eta\tau)\mathsf{KL}\left(\pi_\tau^\star\,\|\,\pi^{(2\gamma)}\right)-\eta\sum_{t=2\gamma+1}^{T}\sum_{i\in V}(\overline{\pi}_i^{(\kappa_i^{(t)})}-\pi_{i,\tau}^\star)^\top A_i(\overline{\pi}^{(\kappa_i^{(t)})}-\pi_\tau^\star)\\
&\quad-\left(1-28\eta d_{\mathsf{max}}\left\|A\right\|_\infty(\gamma+1)^{5/2}\right)\sum_{t=2\gamma+1}^{T}\psi^{(t)}+14\eta d_{\mathsf{max}}\left\|A\right\|_\infty(\gamma+1)^{5/2}\sum_{l\in\Gamma}\psi^{(l)}\\
&\leq(1-\eta\tau)\mathsf{KL}\left(\pi_\tau^\star\,\|\,\pi^{(2\gamma)}\right)-\eta\sum_{t=2\gamma+1}^{T}\sum_{i\in V}(\overline{\pi}_i^{(\kappa_i^{(t)})}-\pi_{i,\tau}^\star)^\top A_i(\overline{\pi}^{(\kappa_i^{(t)})}-\pi_\tau^\star)+\frac{1}{3}\sum_{l\in\Gamma}\psi^{(l)},
\end{aligned}\tag{E.44}$$

where $\Gamma=\{1,\cdots,2\gamma\}\cup\{T+1,\cdots,T+\gamma\}$. The last step results from the choice of learning rate $\eta\leq\frac{1}{28d_{\mathsf{max}}\|A\|_\infty(\gamma+1)^{5/2}}$. It now remains to bound the terms $\sum_{t=2\gamma+1}^{T}\sum_{i\in V}(\overline{\pi}_i^{(\kappa_i^{(t)})}-\pi_{i,\tau}^\star)^\top A_i(\overline{\pi}^{(\kappa_i^{(t)})}-\pi_\tau^\star)$, $\mathsf{KL}\left(\pi_\tau^\star\,\|\,\pi^{(2\gamma)}\right)$ and $\sum_{l\in\Gamma}\psi^{(l)}$. In view of Lemma 35, we have

$$\begin{aligned}
&-\sum_{t=2\gamma+1}^{T}\sum_{i\in V}(\overline{\pi}_i^{(\kappa_i^{(t)})}-\pi_{i,\tau}^\star)^\top A_i(\overline{\pi}^{(\kappa_i^{(t)})}-\pi_\tau^\star)\\
&=\sum_{t=\gamma+1}^{T}\sum_{i\in V}(\overline{\pi}_i^{(t)}-\pi_{i,\tau}^\star)^\top A_i(\overline{\pi}^{(t)}-\pi_\tau^\star)-\sum_{t=2\gamma+1}^{T}\sum_{i\in V}(\overline{\pi}_i^{(\kappa_i^{(t)})}-\pi_{i,\tau}^\star)^\top A_i(\overline{\pi}^{(\kappa_i^{(t)})}-\pi_\tau^\star).
\end{aligned}$$

We remark that each $(\overline{\pi}_i^{(\kappa_i^{(t)})} - \pi_{i,\tau}^\star)^\top A_i(\overline{\pi}^{(\kappa_i^{(t)})} - \pi_\tau^\star)$ term will cancel out due to the mapping $t \mapsto \kappa_i^{(t)}$ being injective when $t \geq \gamma$. In addition, we have a crude bound

$$(\overline{\pi}_i^{(t)} - \pi_{i,\tau}^\star)^\top A_i(\overline{\pi}^{(t)} - \pi_\tau^\star) = \sum_{j \in \mathcal{N}_i} (\overline{\pi}_i^{(t)} - \pi_{i,\tau}^\star)^\top A_{ij}(\overline{\pi}_j^{(t)} - \pi_{j,\tau}^\star) \leq 4 d_{\mathsf{max}} \|A\|_\infty$$

for every $i \in V, t \geq 0$. Applying the bound to the remaining $n\gamma$ terms gives

$$-\sum_{t=2\gamma+1}^{T} \sum_{i \in V} (\overline{\pi}_i^{(\kappa_i^{(t)})} - \pi_{i,\tau}^\star)^\top A_i(\overline{\pi}^{(\kappa_i^{(t)})} - \pi_\tau^\star) \leq 4n\gamma d_{\mathsf{max}} \|A\|_\infty . \tag{E.45}$$

The remaining terms $\mathsf{KL}\left(\pi_\tau^\star \,\|\, \pi^{(2\gamma)}\right)$ and $\psi^{(l)}$ can be bounded with the following lemma, with the proof postponed to Appendix E.3.8.

Lemma 42. *It holds for all $i \in V$ and $t \geq 0$ that*

$$\psi_i^{(t)} \leq \eta(d_{\mathsf{max}} \|A\|_\infty (2\gamma + 11) + 3\tau \log |S_i|). \tag{E.46}$$

In addition, we have

$$\mathsf{KL}\left(\pi_{i,\tau}^\star \,\|\, \pi_i^{(2\gamma)}\right) \leq \mathsf{KL}\left(\pi_{i,\tau}^\star \,\|\, \pi_i^{(0)}\right) + 4\eta d_{\mathsf{max}} \|A\|_\infty \gamma. \tag{E.47}$$

Putting all pieces together, we continue from (E.44) and show that

$$\begin{aligned}
\eta\tau \sum_{t=2\gamma+1}^{T} \mathsf{KL}\left(\pi_\tau^\star \,\|\, \pi^{(t)}\right) &\leq (1-\eta\tau)\mathsf{KL}\left(\pi_\tau^\star \,\|\, \pi^{(2\gamma)}\right) - \eta \sum_{t=2\gamma+1}^{T} \sum_{i \in V} (\overline{\pi}_i^{(\kappa_i^{(t)})} - \pi_{i,\tau}^\star)^\top A_i(\overline{\pi}^{(\kappa_i^{(t)})} - \pi_\tau^\star) + \frac{1}{3}\sum_{l \in \Gamma} \psi^{(l)} \\
&\leq \mathsf{KL}\left(\pi_{i,\tau}^\star \,\|\, \pi_i^{(0)}\right) + 8\eta n\gamma d_{\mathsf{max}} \|A\|_\infty + \eta\gamma\Big(n d_{\mathsf{max}} \|A\|_\infty (2\gamma + 11) + 3\tau \sum_{i \in V} \log |S_i|\Big) \\
&\leq \mathsf{KL}\left(\pi_{i,\tau}^\star \,\|\, \pi_i^{(0)}\right) + 8\eta n\Big[\gamma d_{\mathsf{max}} \|A\|_\infty + \gamma\Big(d_{\mathsf{max}} \|A\|_\infty (2\gamma + 11) + 3\tau \log S_{\max}\Big)\Big] \\
&\leq \mathsf{KL}\left(\pi_{i,\tau}^\star \,\|\, \pi_i^{(0)}\right) + n + 24\eta\tau n\gamma \log S_{\max}.
\end{aligned}$$

Bounding the term $\mathsf{KL}\left(\pi_\tau^\star \,\|\, \overline{\pi}^{(t-\gamma+1)}\right)$. By definition of KL divergence, we have

$$\begin{aligned}
&\mathsf{KL}\left(\pi_{i,\tau}^\star \,\|\, \overline{\pi}_i^{(t-\gamma+1)}\right) \\
&= \mathsf{KL}\left(\pi_{i,\tau}^\star \,\|\, \pi_i^{(t+1)}\right) + \left\langle \pi_{i,\tau}^\star, \log \pi_i^{(t+1)} - \log \overline{\pi}_i^{(t-\gamma+1)} \right\rangle \\
&= \mathsf{KL}\left(\pi_{i,\tau}^\star \,\|\, \pi_i^{(t+1)}\right) - \mathsf{KL}\left(\overline{\pi}_i^{(t-\gamma+1)} \,\|\, \pi_i^{(t+1)}\right) + \left\langle \pi_{i,\tau}^\star - \overline{\pi}_i^{(t-\gamma+1)}, \log \pi_i^{(t+1)} - \log \overline{\pi}_i^{(t-\gamma+1)} \right\rangle .
\end{aligned} \tag{E.48}$$

It follows directly from the update rules that

$$\begin{cases}
\log \overline{\pi}_i^{(t-\gamma+1)} \overset{\mathbf{1}}{=} (1-\overline{\eta}\tau) \log \pi_i^{(t-\gamma)} + \overline{\eta} A_i \overline{\pi}^{(\kappa_i^{(t-\gamma)})} \\
\log \pi_i^{(t+1)} \overset{\mathbf{1}}{=} (1-\eta\tau)^{\gamma+1} \log \pi_i^{(t-\gamma)} + \eta \sum_{l=t-\gamma+1}^{t+1} (1-\eta\tau)^{t-l+1} A_i \overline{\pi}^{(\kappa_i^{(l)})},
\end{cases}$$

which enables us to control the term $\left\langle \pi^\star_{i,\tau} - \overline{\pi}_i^{(t-\gamma+1)}, \log \pi_i^{(t+1)} - \log \overline{\pi}_i^{(t-\gamma+1)} \right\rangle$ as

$$\begin{aligned}
&\left\langle \pi^\star_{i,\tau} - \overline{\pi}_i^{(t-\gamma+1)}, \log \pi_i^{(t+1)} - \log \overline{\pi}_i^{(t-\gamma+1)} \right\rangle \\
&= \eta \sum_{l=t-\gamma+1}^{t+1} (1-\eta\tau)^{t-l+1} \left\langle \pi^\star_{i,\tau} - \overline{\pi}_i^{(t-\gamma+1)},\, A_i(\overline{\pi}^{(\kappa_i^{(t-\gamma)})} - \overline{\pi}^{(\kappa_i^{(l)})}) \right\rangle \\
&\le \eta \left\|A\right\|_\infty \left\|\pi^\star_{i,\tau} - \overline{\pi}_i^{(t-\gamma+1)}\right\|_1 \sum_{j\in\mathcal{N}_i} \sum_{l=t-\gamma+1}^{t+1} \left\|\overline{\pi}_j^{(\kappa_i^{(t-\gamma)})} - \overline{\pi}_j^{(\kappa_i^{(l)})}\right\|_1. \qquad \text{(E.49)}
\end{aligned}$$

In the same vein as Lemma 41, we can bound the term $\sum_{l=t-\gamma+1}^{t+1} \left\|\overline{\pi}_j^{(\kappa_i^{(t-\gamma)})} - \overline{\pi}_j^{(\kappa_i^{(l)})}\right\|_1$ with $\{\psi_i^{(l)}\}$, as detailed in the following lemma. The proof is omitted due to its similarity with that of Lemma 41.

Lemma 43. *When $t \ge 2\gamma$, it holds that*

$$\sum_{l=t-\gamma+1}^{t+1} \left\|\overline{\pi}_j^{(\kappa_i^{(t-\gamma)})} - \overline{\pi}_j^{(\kappa_i^{(l)})}\right\|_1 \le 4\sqrt{2}(\gamma+1) \sum_{l=t-2\gamma+1}^{t+\gamma+1} \sqrt{\psi_i^{(l)}} + 2\sqrt{2}(\gamma+1)^2 \sqrt{\psi_j^{(\nu_j(\kappa_i^{(t-\gamma)}))}}.$$

Plugging the above lemma into (E.49), we have

$$\begin{aligned}
&\left\langle \pi^\star_{i,\tau} - \overline{\pi}_i^{(t-\gamma+1)}, \log \pi_i^{(t+1)} - \log \overline{\pi}_i^{(t-\gamma+1)} \right\rangle \\
&\le \eta \left\|A\right\|_\infty \left\|\pi^\star_{i,\tau} - \overline{\pi}_i^{(t-\gamma+1)}\right\|_1 \sum_{j\in\mathcal{N}_i} \left(4\sqrt{2}(\gamma+1) \sum_{l=t-2\gamma+1}^{t+\gamma+1} \sqrt{\psi_i^{(l)}} + 2\sqrt{2}(\gamma+1)^2 \sqrt{\psi_j^{(\nu_j(\kappa_i^{(t-\gamma)}))}} \right) \\
&\overset{\text{(i)}}{\le} \frac{1}{2}\eta \left\|A\right\|_\infty \Bigg\{ 14 d_{\mathsf{max}} (\gamma+1)^{3/2} \left\|\pi^\star_{i,\tau} - \overline{\pi}_i^{(t-\gamma+1)}\right\|_1^2 \\
&\qquad + \sum_{j\in\mathcal{N}_i} \left[8(\gamma+1)^{3/2} \sum_{l=t-2\gamma+1}^{t+\gamma+1} \psi_j^{(l)} + 4(\gamma+1)^{5/2} \psi_j^{(\nu_j(\kappa_i^{(t-\gamma)}))} \right] \Bigg\} \\
&\overset{\text{(ii)}}{\le} \eta \left\|A\right\|_\infty \Bigg\{ 14 d_{\mathsf{max}} (\gamma+1)^{3/2} \mathsf{KL}\left(\pi^\star_{i,\tau} \,\|\, \overline{\pi}_i^{(t-\gamma+1)}\right) \\
&\qquad + \sum_{j\in\mathcal{N}_i} \left[4(\gamma+1)^{3/2} \sum_{l=t-2\gamma+1}^{t+\gamma+1} \psi_j^{(l)} + 2(\gamma+1)^{5/2} \psi_j^{(\nu_j(\kappa_i^{(t-\gamma)}))} \right] \Bigg\},
\end{aligned}$$

where (i) results from similar arguments in (E.42) and (ii) invokes Pinsker's inequality. Substitution of the above inequality into (E.48) and summing over $i \in V$ leads to

$$\begin{aligned}
&(1 - 14\eta d_{\mathsf{max}} \left\|A\right\|_\infty (\gamma+1)^{3/2}) \mathsf{KL}\left(\pi^\star_\tau \,\|\, \overline{\pi}^{(t-\gamma+1)}\right) \\
&\le \mathsf{KL}\left(\pi^\star_\tau \,\|\, \pi^{(t+1)}\right) + \eta \left\|A\right\|_\infty \sum_{(i,j)\in E} \left[4(\gamma+1)^{3/2} \sum_{l=t-2\gamma+1}^{t+\gamma+1} \psi_j^{(l)} + 2(\gamma+1)^{5/2} \psi_j^{(\nu_j(\kappa_i^{(t-\gamma)}))} \right] \\
&\le \mathsf{KL}\left(\pi^\star_\tau \,\|\, \pi^{(t+1)}\right) + 4\eta d_{\mathsf{max}} \left\|A\right\|_\infty (\gamma+1)^{3/2} \sum_{l=t-2\gamma+1}^{t+\gamma+1} \psi^{(l)} + 2\eta d_{\mathsf{max}} \left\|A\right\|_\infty (\gamma+1)^{5/2} \psi^{(\nu_j(\kappa_i^{(t-\gamma)}))}.
\end{aligned}$$

Summing the above inequality over $t = 2\gamma-1, \cdots, T-1$ and adding $\sum_{t=2\gamma}^{T-1}\sum_{i\in V}\mathsf{KL}\Big(\overline{\pi}_i^{(\kappa_i^{(t+1)})} \,\|\, \pi_{i,\tau}^\star\Big)$ to the both sides,

$$
\begin{aligned}
&\sum_{t=2\gamma}^{T-1}\left[\frac{2}{3}\mathsf{KL}\left(\pi_\tau^\star \,\|\, \overline{\pi}^{(t-\gamma+1)}\right) + \sum_{i\in V}\mathsf{KL}\Big(\overline{\pi}_i^{(\kappa_i^{(t+1)})} \,\|\, \pi_{i,\tau}^\star\Big)\right] \\
&\le \sum_{t=2\gamma}^{T-1}\mathsf{KL}\left(\pi_\tau^\star \,\|\, \pi^{(t+1)}\right) + \sum_{t=2\gamma}^{T-1}\sum_{i\in V}\mathsf{KL}\Big(\overline{\pi}_i^{(\kappa_i^{(t+1)})} \,\|\, \pi_{i,\tau}^\star\Big) \\
&\quad + 4\eta d_{\mathsf{max}}\left\|A\right\|_\infty(\gamma+1)^{3/2}\sum_{t=2\gamma}^{T-1}\sum_{l=t-2\gamma+1}^{t+\gamma+1}\psi^{(l)} + 2\eta d_{\mathsf{max}}\left\|A\right\|_\infty(\gamma+1)^{5/2}\sum_{t=2\gamma}^{T-1}\psi^{(\nu_j(\kappa_i^{(t-\gamma)}))} \\
&\overset{\text{(i)}}{\le} \frac{1}{\eta\tau}\Bigg\{(1-\eta\tau)\mathsf{KL}\left(\pi_\tau^\star \,\|\, \pi^{(2\gamma)}\right) - \eta\sum_{t=2\gamma+1}^{T}\sum_{i\in V}(\overline{\pi}_i^{(\kappa_i^{(t)})} - \pi_{i,\tau}^\star)^\top A_i(\overline{\pi}^{(\kappa_i^{(t)})} - \pi_\tau^\star) \\
&\quad - \Big(1 - 28\eta d_{\mathsf{max}}\left\|A\right\|_\infty(\gamma+1)^{5/2}\Big)\sum_{t=2\gamma+1}^{T}\psi^{(t)} + 14\eta d_{\mathsf{max}}\left\|A\right\|_\infty(\gamma+1)^{5/2}\sum_{l\in\Gamma}\psi^{(l)}\Bigg\} \\
&\quad + 12\eta d_{\mathsf{max}}\left\|A\right\|_\infty(\gamma+1)^{5/2}\sum_{l=1}^{T+\gamma}\psi^{(l)} + 2\eta d_{\mathsf{max}}\left\|A\right\|_\infty(\gamma+1)^{5/2}\sum_{t=0}^{T+\gamma-1}\psi^{(l)} \\
&= \frac{1}{\eta\tau}\Bigg\{(1-\eta\tau)\mathsf{KL}\left(\pi_\tau^\star \,\|\, \pi^{(2\gamma)}\right) - \eta\sum_{t=2\gamma+1}^{T}\sum_{i\in V}(\overline{\pi}_i^{(\kappa_i^{(t)})} - \pi_{i,\tau}^\star)^\top A_i(\overline{\pi}^{(\kappa_i^{(t)})} - \pi_\tau^\star) \\
&\quad - \Big(1 - 28(1+\frac{\eta\tau}{2})\eta d_{\mathsf{max}}\left\|A\right\|_\infty(\gamma+1)^{5/2}\Big)\sum_{t=2\gamma+1}^{T}\psi^{(t)} + 14(1+\eta\tau)\eta d_{\mathsf{max}}\left\|A\right\|_\infty(\gamma+1)^{5/2}\sum_{l\in\Gamma}\psi^{(l)}\Bigg\}.
\end{aligned}
$$

Here, (i) invokes the bound established in (E.44). We remark that our choice of learning rate

$$
\eta \le \min\left\{\frac{1}{2\tau(\gamma+1)}, \frac{1}{42 d_{\mathsf{max}}\left\|A\right\|_\infty(\gamma+1)^{5/2}}\right\}
$$

guarantees $1 - 28(1+\frac{\eta\tau}{2})\eta d_{\mathsf{max}}\left\|A\right\|_\infty(\gamma+1)^{5/2} \ge 0$. This taken together with (E.45) and Lemma 42 gives

$$
\begin{aligned}
&\sum_{t=2\gamma}^{T-1}\left[\frac{2}{3}\mathsf{KL}\left(\pi_\tau^\star \,\|\, \overline{\pi}^{(t-\gamma+1)}\right) + \sum_{i\in V}\mathsf{KL}\Big(\overline{\pi}_i^{(\kappa_i^{(t+1)})} \,\|\, \pi_{i,\tau}^\star\Big)\right] \\
&\le \frac{1}{\eta\tau}\Bigg\{(1-\eta\tau)\mathsf{KL}\left(\pi_\tau^\star \,\|\, \pi^{(2\gamma)}\right) - \eta\sum_{t=2\gamma+1}^{T}\sum_{i\in V}(\overline{\pi}_i^{(\kappa_i^{(t)})} - \pi_{i,\tau}^\star)^\top A_i(\overline{\pi}^{(\kappa_i^{(t)})} - \pi_\tau^\star) + \frac{1}{2}\sum_{l\in\Gamma}\psi^{(l)}\Bigg\} \\
&\le \frac{1}{\eta\tau}\left\{\mathsf{KL}\left(\pi_{i,\tau}^\star \,\|\, \pi_i^{(0)}\right) + 8\eta n\Big[\gamma d_{\mathsf{max}}\left\|A\right\|_\infty + \frac{3\gamma}{2}\Big(d_{\mathsf{max}}\left\|A\right\|_\infty(2\gamma+11) + 3\tau\log S_{\max}\Big)\Big]\right\} \\
&\le \frac{1}{\eta\tau}\left\{\mathsf{KL}\left(\pi_{i,\tau}^\star \,\|\, \pi_i^{(0)}\right) + n + 36\eta\tau n\gamma\log S_{\max}\right\}. \qquad \text{(E.50)}
\end{aligned}
$$

Bounding the QRE gap. With Lemma 37, we have

$$\sum_{t=2\gamma}^{T-\gamma-1} \texttt{QRE-Gap}_\tau(\overline{\pi}^{(t+1)}) \le \sum_{t=2\gamma}^{T-\gamma-1} \left(\frac{d_{\max}^2 \|A\|_\infty^2}{\tau} \mathsf{KL}\left(\pi_\tau^\star \,\|\, \overline{\pi}^{(t+1)}\right) + \tau \mathsf{KL}\left(\overline{\pi}^{(t+1)} \,\|\, \pi_\tau^\star\right) \right)$$
$$\le \max\left\{ \frac{3d_{\max}^2 \|A\|_\infty^2}{2\tau}, \tau \right\} \sum_{t=2\gamma}^{T-\gamma-1} \left(\frac{2}{3} \mathsf{KL}\left(\pi_\tau^\star \,\|\, \overline{\pi}^{(t+1)}\right) + \mathsf{KL}\left(\overline{\pi}^{(t+1)} \,\|\, \pi_\tau^\star\right) \right).$$

Since the mapping $t \mapsto \nu_i(t)$ is injective, we have

$$\sum_{t=2\gamma}^{T-\gamma-1} \sum_{i\in V} \mathsf{KL}\left(\overline{\pi}_i^{(t+1)} \,\|\, \pi_{i,\tau}^\star\right) = \sum_{t=2\gamma}^{T-\gamma-1} \sum_{i\in V} \mathsf{KL}\left(\overline{\pi}_i^{(\kappa_i^{(\nu_i(t+1))})} \,\|\, \pi_{i,\tau}^\star\right) \le \sum_{t=2\gamma}^{T-1} \sum_{i\in V} \mathsf{KL}\left(\overline{\pi}_i^{(\kappa_i^{(t+1)})} \,\|\, \pi_{i,\tau}^\star\right).$$

Combining the above two equalities gives

$$\sum_{t=2\gamma}^{T-\gamma-1} \texttt{QRE-Gap}_\tau(\overline{\pi}^{(t+1)}) \le \max\left\{ \frac{3d_{\max}^2 \|A\|_\infty^2}{2\tau}, \tau \right\} \left[\sum_{t=2\gamma}^{T-\gamma-1} \frac{2}{3} \mathsf{KL}\left(\pi_\tau^\star \,\|\, \overline{\pi}^{(t+1)}\right) + \sum_{t=2\gamma}^{T-1} \mathsf{KL}\left(\overline{\pi}^{(t+1)} \,\|\, \pi_\tau^\star\right) \right]$$
$$\le \max\left\{ \frac{3d_{\max}^2 \|A\|_\infty^2}{2\tau}, \tau \right\} \sum_{t=2\gamma}^{T-1} \left[\frac{2}{3} \mathsf{KL}\left(\pi_\tau^\star \,\|\, \overline{\pi}^{(t-\gamma+1)}\right) + \sum_{i\in V} \mathsf{KL}\left(\overline{\pi}_i^{(\kappa_i^{(t+1)})} \,\|\, \pi_{i,\tau}^\star\right) \right]$$
$$\le \max\left\{ \frac{3d_{\max}^2 \|A\|_\infty^2}{2\tau}, \tau \right\} \frac{1}{\eta\tau} \left(\mathsf{KL}\left(\pi_{i,\tau}^\star \,\|\, \pi_i^{(0)}\right) + n + 36\eta\tau n\gamma \log S_{\max} \right),$$

where the last step results from (E.50).

E.3 Proof of auxiliary lemmas

E.3.1 Proof of Lemma 35

To prove this lemma, we recall a key observation in Cai et al. [2016] that allows one to transform a zero-sum polymatrix game $\mathcal{G} = \{(V, E), \{S_i\}_{i\in V}, \{A_{ij}\}_{(i,j)\in E}\}$ into a pairwise constant-sum polymatrix game $\widetilde{\mathcal{G}} = \{(V, E), \{S_i\}_{i\in V}, \{\widetilde{A}_{ij}\}_{(i,j)\in E}\}$ such that

(1) For every player $i \in V$, it has the same payoff in $\mathcal{G}$ and $\widetilde{\mathcal{G}}$:

$$u_i(s) = \widetilde{u}_i(s), \qquad \forall s \in S.$$

(2) For each pair $(i, j) \in E$, $i \neq j$, the two-player game $\widetilde{\mathcal{G}}$ is constant-sum, i.e., there exist constants $\alpha_{ij} = \alpha_{ji}$, such that

$$\widetilde{A}_{ij}(s_i, s_j) + \widetilde{A}_{ji}(s_j, s_i) = \alpha_{ij} \tag{E.51}$$

holds for all $s_i \in S_i, s_j \in S_j$.

We are now in a place to prove Lemma 35. Let $\widetilde{\mathcal{G}}$ be the pairwise constant-sum polymatrix game associated with $\mathcal{G}$ after the above payoff preserving transformation. We have

$$\begin{aligned}\sum_{i\in V}\big[u_i(\pi_i,\pi'_{-i})+u_i(\pi'_i,\pi_{-i})\big] &= \sum_{i\in V}\big[\widetilde{u}_i(\pi_i,\pi'_{-i})+\widetilde{u}_i(\pi'_i,\pi_{-i})\big]\\ &= \sum_{(i,j)\in E}\left[\mathop{\mathbb{E}}_{s_i\sim\pi_i,s_j\sim\pi'_j}\Big[\widetilde{A}_{ij}(s_i,s_j)\Big]+\mathop{\mathbb{E}}_{s_i\sim\pi'_i,s_j\sim\pi_j}\Big[\widetilde{A}_{ij}(s_i,s_j)\Big]\right]\\ &= \sum_{(i,j)\in E}\left[\mathop{\mathbb{E}}_{s_i\sim\pi_i,s_j\sim\pi'_j}\Big[\widetilde{A}_{ij}(s_i,s_j)\Big]+\mathop{\mathbb{E}}_{s_i\sim\pi'_i,s_j\sim\pi_j}\Big[\alpha_{ij}-\widetilde{A}_{ji}(s_j,s_i)\Big]\right]\\ &= \sum_{(i,j)\in E}\alpha_{ij}=0,\end{aligned}$$

where the penultimate line uses (E.51), and the last line uses the fact that $\widetilde{\mathcal{G}}$ is also a zero-sum polymatrix game, which satisfies

$$\sum_{(i,j)\in E}\alpha_{ij}=\sum_{(i,j)\in E}\Big[\widetilde{A}_{ij}(s_i,s_j)+\widetilde{A}_{ji}(s_j,s_i)\Big]=\sum_{i\in V}\widetilde{u}_i(s)+\sum_{j\in V}\widetilde{u}_j(s)=0$$

for any arbitrary $s\in S$.

E.3.2 Proof of Lemma 36

In view of the update rule (6.9), we have

$$\log\pi_i^{(t+1)}=(1-\eta\tau)\log\pi_i^{(t)}+\eta A_i\overline{\pi}^{(t+1)}+c_i\mathbf{1}$$

for some constant c_i. On the other hand, it follows from the expression of QRE in (6.5) that

$$\eta\tau\log\pi_{i,\tau}^{\star}=\eta A_i\pi_\tau^{\star}+c_i^{\star}\mathbf{1} \tag{E.52}$$

for some constant $c_i^{\star}$. By combining the above two equalities and taking the inner product with $\overline{\pi}_i^{(t+1)}-\pi_{i,\tau}^{\star}$, we have

$$\Big\langle\log\pi_i^{(t+1)}-(1-\eta\tau)\log\pi_i^{(t)}-\eta\tau\log\pi_{i,\tau}^{\star},\,\overline{\pi}_i^{(t+1)}-\pi_{i,\tau}^{\star}\Big\rangle=\eta(\overline{\pi}_i^{(t+1)}-\pi_{i,\tau}^{\star})^\top A_i(\overline{\pi}^{(t+1)}-\pi_\tau^{\star}). \tag{E.53}$$

Summing the above equality over $i\in V$ gives

$$\begin{aligned}&\big\langle\log\pi^{(t+1)}-(1-\eta\tau)\log\pi^{(t)}-\eta\tau\log\pi_\tau^{\star},\,\overline{\pi}^{(t+1)}-\pi_\tau^{\star}\big\rangle\\ &=\eta\sum_{i\in V}(\overline{\pi}_i^{(t+1)}-\pi_{i,\tau}^{\star})^\top A_i(\overline{\pi}^{(t+1)}-\pi_\tau^{\star})\\ &=\eta\sum_{i\in V}\Big[(\overline{\pi}_i^{(t+1)})^\top A_i\overline{\pi}^{(t+1)}+(\pi_{i,\tau}^{\star})^\top A_i\pi_\tau^{\star}\Big]-\eta\sum_{i\in V}\Big[(\overline{\pi}_i^{(t+1)})^\top A_i\pi_\tau^{\star}+(\pi_{i,\tau}^{\star})^\top A_i\overline{\pi}^{(t+1)}\Big]\\ &=\eta\sum_{i\in V}\Big[u_i(\overline{\pi}^{(t+1)})+u_i(\pi_\tau^{\star})\Big]=0,\end{aligned}$$

where the last line follows from $\sum_{i\in V}\Big[(\overline{\pi}_i^{(t+1)})^\top A_i\pi_\tau^{\star}+(\pi_{i,\tau}^{\star})^\top A_i\overline{\pi}^{(t+1)}\Big]=0$ due to Lemma 35, as well as that the game is zero-sum.

E.3.3 Proof of Lemma 37

Recalling the definition

$$\begin{aligned}\texttt{QRE-Gap}_\tau(\pi) &= \max_{i\in V}\left[\max_{\pi_i'\in\Delta(S_i)} u_{i,\tau}(\pi_i',\pi_{-i}) - u_{i,\tau}(\pi)\right]\\ &\le \sum_{i\in V}\left[\max_{\pi_i'\in\Delta(S_i)} u_{i,\tau}(\pi_i',\pi_{-i}) - u_{i,\tau}(\pi)\right]\\ &= \max_{i\in V:\pi_i'\in\Delta(S_i)}\sum_{i\in V}\left[u_{i,\tau}(\pi_i',\pi_{-i}) - u_{i,\tau}(\pi_i,\pi_{-i})\right],\end{aligned}$$

where the inequality holds since $\max_{\pi_i'\in\Delta(S_i)} u_{i,\tau}(\pi_i',\pi_{-i}) - u_{i,\tau}(\pi) \ge u_{i,\tau}(\pi_i,\pi_{-i}) - u_{i,\tau}(\pi) = 0$ for all $i\in V$. We now proceed to decompose

$$\begin{aligned}&\sum_{i\in V}\left[u_{i,\tau}(\pi_i',\pi_{-i}) - u_{i,\tau}(\pi_i,\pi_{-i})\right]\\ &= \sum_{i\in V}\left[u_{i,\tau}(\pi_i',\pi_{-i}) - u_{i,\tau}(\pi_{i,\tau}^\star,\pi_{-i,\tau}^\star)\right] - \tau\sum_{i\in V}\left(\mathcal{H}(\pi_i) - \mathcal{H}(\pi_{i,\tau}^\star)\right)\\ &= \sum_{i\in V}\left[u_{i,\tau}(\pi_i',\pi_{-i}) - u_{i,\tau}(\pi_i',\pi_{-i,\tau}^\star) - u_{i,\tau}(\pi_{i,\tau}^\star,\pi_{-i}) + u_{i,\tau}(\pi_{i,\tau}^\star,\pi_{-i,\tau}^\star)\right]\\ &\quad + \sum_{i\in V}\left[u_{i,\tau}(\pi_{i,\tau}^\star,\pi_{-i}) - u_{i,\tau}(\pi_{i,\tau}^\star,\pi_{-i,\tau}^\star) - \tau\left(\mathcal{H}(\pi_i) - \mathcal{H}(\pi_{i,\tau}^\star)\right)\right]\\ &\quad + \sum_{i\in V}\left[u_{i,\tau}(\pi_i',\pi_{-i,\tau}^\star) - u_{i,\tau}(\pi_{i,\tau}^\star,\pi_{-i,\tau}^\star)\right] \qquad \text{(E.54)}\end{aligned}$$

where the first line follows from $\sum_{i\in V}\left(u_{i,\tau}(\pi) - \tau\mathcal{H}(\pi_i)\right) = \sum_{i\in V}\left(u_{i,\tau}(\pi_\tau^\star) - \tau\mathcal{H}(\pi_{i,\tau}^\star)\right) = 0$ by the definition of zero-sum games. It boils down to control the terms on the RHS of (E.54).

- To control the first term, by the definition of $u_{i,\tau}$ in (6.6) (see also (6.3)), it follows that

$$\begin{aligned}&u_{i,\tau}(\pi_i',\pi_{-i}) - u_{i,\tau}(\pi_i',\pi_{-i,\tau}^\star) - u_{i,\tau}(\pi_{i,\tau}^\star,\pi_{-i}) + u_{i,\tau}(\pi_{i,\tau}^\star,\pi_{-i,\tau}^\star)\\ &= u_i(\pi_i',\pi_{-i}) - u_i(\pi_i',\pi_{-i,\tau}^\star) - u_i(\pi_{i,\tau}^\star,\pi_{-i}) + u_i(\pi_{i,\tau}^\star,\pi_{-i,\tau}^\star)\\ &= (\pi_i'-\pi_{i,\tau}^\star)^\top A_i(\pi-\pi_\tau^\star) = \sum_{j\in\mathcal{N}_i}(\pi_i'-\pi_{i,\tau}^\star)^\top A_{ij}(\pi_j-\pi_{j,\tau}^\star),\end{aligned}$$

  which each summand can be further bounded by Young's inequality and Pinsker's inequality as

$$\begin{aligned}(\pi_i'-\pi_{i,\tau}^\star)^\top A_{ij}(\pi_j-\pi_{j,\tau}^\star) &\le \|A\|_\infty\left\|\pi_i'-\pi_{i,\tau}^\star\right\|_1\left\|\pi_j-\pi_{j,\tau}^\star\right\|_1\\ &\le \frac{1}{2}\|A\|_\infty\left(\frac{\tau}{d_{\max}\|A\|_\infty}\left\|\pi_i'-\pi_{i,\tau}^\star\right\|_1^2 + \frac{d_{\max}\|A\|_\infty}{\tau}\left\|\pi_j-\pi_{j,\tau}^\star\right\|_1^2\right)\\ &\le \|A\|_\infty\left(\frac{\tau}{d_{\max}\|A\|_\infty}\mathsf{KL}\left(\pi_i'\,\|\,\pi_{i,\tau}^\star\right) + \frac{d_{\max}\|A\|_\infty}{\tau}\mathsf{KL}\left(\pi_{j,\tau}^\star\,\|\,\pi_j\right)\right).\end{aligned}$$

  Summing the inequality over i,j gives

$$\begin{aligned}&\sum_{i\in V}\left[u_{i,\tau}(\pi_i',\pi_{-i}) - u_{i,\tau}(\pi_i',\pi_{-i,\tau}^\star) - u_{i,\tau}(\pi_{i,\tau}^\star,\pi_{-i}) + u_{i,\tau}(\pi_{i,\tau}^\star,\pi_{-i,\tau}^\star)\right]\\ &\le \tau\mathsf{KL}\left(\pi'\,\|\,\pi_\tau^\star\right) + \frac{d_{\max}^2\|A\|_\infty^2}{\tau}\mathsf{KL}\left(\pi_\tau^\star\,\|\,\pi\right). \qquad \text{(E.55)}\end{aligned}$$

- Regarding the second term, we have

$$\begin{aligned}
&\sum_{i\in V}\left[u_{i,\tau}(\pi^\star_{i,\tau},\pi_{-i})-u_{i,\tau}(\pi^\star_{i,\tau},\pi^\star_{-i,\tau})-\tau\left(\mathcal{H}(\pi_i)-\mathcal{H}(\pi^\star_{i,\tau})\right)\right]\\
&=\sum_{i\in V}\left[(\pi^\star_{i,\tau})^\top A_i(\pi-\pi^\star_\tau)+\tau(\pi_i^\top\log\pi_i-(\pi^\star_{i,\tau})^\top\log\pi^\star_{i,\tau})\right]\\
&=\sum_{i\in V}\left[(\pi^\star_{i,\tau})^\top A_i(\pi-\pi^\star_\tau)+\tau\left(\langle\pi_i,\log\pi_i-\log\pi^\star_{i,\tau}\rangle+\langle\pi_i-\pi^\star_{i,\tau},\log\pi^\star_{i,\tau}\rangle\right)\right]\\
&=\sum_{i\in V}\left[(\pi^\star_{i,\tau})^\top A_i(\pi-\pi^\star_\tau)+(\pi_i-\pi^\star_{i,\tau})^\top A_i\pi^\star_\tau+\tau\mathsf{KL}\left(\pi_i\,\|\,\pi^\star_{i,\tau}\right)\right]\\
&=\tau\mathsf{KL}\left(\pi\,\|\,\pi^\star_\tau\right),
\end{aligned}\tag{E.56}$$

  where the penultimate step follows from (E.52) and the last step invokes Lemma 35.

- Moving to the last term, we have

$$\begin{aligned}
u_{i,\tau}(\pi^\star_{i,\tau},\pi^\star_{-i,\tau})-u_{i,\tau}(\pi'_i,\pi^\star_{-i,\tau})&=(\pi^\star_{i,\tau}-\pi'_i)^\top A_i\pi^\star_\tau-\tau(\pi^\star_{i,\tau})^\top\log\pi^\star_{i,\tau}+\tau(\pi'_i)^\top\log\pi'_i\\
&=\tau(\pi^\star_{i,\tau}-\pi'_i)^\top\log\pi^\star_{i,\tau}-\tau(\pi^\star_{i,\tau})^\top\log\pi^\star_{i,\tau}+\tau(\pi'_i)^\top\log\pi'_i\\
&=\tau\mathsf{KL}\left(\pi'_i\,\|\,\pi^\star_{i,\tau}\right).
\end{aligned}\tag{E.57}$$

  where the second line follows again from (E.52).

Plugging (E.55), (E.56) and (E.57) into (E.54) gives

$$\sum_{i\in V}\left[u_{i,\tau}(\pi'_i,\pi_{-i})-u_{i,\tau}(\pi_i,\pi_{-i})\right]\le\tau\mathsf{KL}\left(\pi\,\|\,\pi^\star_\tau\right)+\frac{d_{\mathsf{max}}^2\|A\|_\infty^2}{\tau}\mathsf{KL}\left(\pi^\star_\tau\,\|\,\pi\right).$$

Taking maximum over π' finishes the proof.

E.3.4 Proof of Lemma 38

Taking logarithm on the both sides of (6.9), we have

$$\log\pi_i^{(t+1)}\overset{\mathbf{1}}{=}(1-\eta\tau)\log\pi_i^{(t)}+\eta A_i\overline{\pi}^{(t-\gamma+1)}.\tag{E.58}$$

On the other hand, the definition of QRE in (6.5) gives

$$\eta\tau\log\pi^\star_{i,\tau}\overset{\mathbf{1}}{=}\eta A_i\pi^\star_\tau.$$

Subtracting the two equalities and taking inner product with $\overline{\pi}_i^{(t-\gamma+1)}-\pi^\star_{i,\tau}$, we get

$$\begin{aligned}
&\left\langle\log\pi_i^{(t+1)}-(1-\eta\tau)\log\pi_i^{(t)}-\eta\tau\log\pi^\star_{i,\tau},\,\overline{\pi}_i^{(t-\gamma+1)}-\pi^\star_{i,\tau}\right\rangle\\
&=\eta\left(\overline{\pi}_i^{(t-\gamma+1)}-\pi^\star_{i,\tau}\right)^\top A_i\left(\overline{\pi}^{(t-\gamma+1)}-\pi^\star_\tau\right).
\end{aligned}$$

Summing the above equality over $i\in V$ leads to

$$\begin{aligned}
&\left\langle\log\pi^{(t+1)}-(1-\eta\tau)\log\pi^{(t)}-\eta\tau\log\pi^\star_\tau,\overline{\pi}_i^{(t-\gamma+1)}-\pi^\star_\tau\right\rangle\\
&=\eta\sum_{i\in V}\left(\overline{\pi}_i^{(t-\gamma+1)}-\pi^\star_{i,\tau}\right)^\top A_i\left(\overline{\pi}^{(t-\gamma+1)}-\pi^\star_\tau\right)=0,
\end{aligned}$$

where the final step results from Lemma 35.

E.3.5 Proof of Lemma 39

Recall from (E.31) that

$$\mathsf{KL}\left(\pi_\tau^\star \,\|\, \pi^{(t+1)}\right) = (1-\eta\tau)\mathsf{KL}\left(\pi_\tau^\star \,\|\, \pi^{(t)}\right) - (1-\eta\tau)\mathsf{KL}\left(\overline{\pi}^{(t-\gamma+1)} \,\|\, \pi^{(t)}\right) - \mathsf{KL}\left(\pi^{(t+1)} \,\|\, \overline{\pi}^{(t-\gamma+1)}\right) + \left\langle \log \overline{\pi}^{(t-\gamma+1)} - \log \pi^{(t+1)}, \overline{\pi}^{(t-\gamma+1)} - \pi^{(t+1)} \right\rangle - \eta\tau\mathsf{KL}\left(\overline{\pi}^{(t-\gamma+1)} \,\|\, \pi_\tau^\star\right). \tag{E.59}$$

When $t < \gamma$, we have $\overline{\pi}_i^{(t-\gamma+1)} = \pi^{(0)}$. It follows that

$$\log \overline{\pi}_i^{(t-\gamma+1)} = \log \pi^{(0)} \overset{\mathbf{1}}{=} 0,$$

and that

$$\begin{aligned}\log \pi_i^{(t+1)} &\overset{\mathbf{1}}{=} (1-\eta\tau)^{t+1} \log \pi^{(0)} + \eta \sum_{l=0}^{t} (1-\eta\tau)^l A_i \overline{\pi}^{(t-\gamma-l+1)} \\ &\overset{\mathbf{1}}{=} \eta \sum_{l=0}^{t} (1-\eta\tau)^l A_i \pi^{(0)}.\end{aligned}$$

Therefore, we can bound the term $\left\langle \log \overline{\pi}^{(t-\gamma+1)} - \log \pi^{(t+1)}, \overline{\pi}^{(t-\gamma+1)} - \pi^{(t+1)} \right\rangle$ as

$$\begin{aligned}\left\langle \log \overline{\pi}^{(t-\gamma+1)} - \log \pi^{(t+1)}, \overline{\pi}^{(t-\gamma+1)} - \pi^{(t+1)} \right\rangle &= \left\langle \eta \sum_{l=0}^{t} (1-\eta\tau)^l A_i \pi^{(0)}, \pi^{(0)} - \pi^{(t+1)} \right\rangle \\ &\le \eta(t+1) d_{\mathsf{max}} \|A\|_\infty \left\| \pi^{(0)} - \pi^{(t+1)} \right\|_1 \\ &\le 2\eta(t+1) d_{\mathsf{max}} \|A\|_\infty. \end{aligned} \tag{E.60}$$

Plugging the above inequality into (E.59) leads to

$$\mathsf{KL}\left(\pi_\tau^\star \,\|\, \pi^{(t+1)}\right) \le (1-\eta\tau)\mathsf{KL}\left(\pi_\tau^\star \,\|\, \pi^{(t)}\right) - (1-\eta\tau)\mathsf{KL}\left(\overline{\pi}^{(t-\gamma+1)} \,\|\, \pi^{(t)}\right) - \mathsf{KL}\left(\pi^{(t+1)} \,\|\, \overline{\pi}^{(t-\gamma+1)}\right) + 2\eta(t+1) d_{\mathsf{max}} \|A\|_\infty.$$

Applying the above inequality recursively to the iterates $0, 1, \ldots, \gamma-1$, we arrive at

$$\begin{aligned}&\mathsf{KL}\left(\pi_\tau^\star \,\|\, \pi^{(\gamma)}\right) \\ &\le (1-\eta\tau)^\gamma \mathsf{KL}\left(\pi_\tau^\star \,\|\, \pi^{(0)}\right) - \sum_{l_1=0}^{\gamma-1} (1-\eta\tau)^{\gamma-1-l_1} \left[(1-\eta\tau)\mathsf{KL}\left(\overline{\pi}^{(l_1-\gamma+1)} \,\|\, \pi^{(l_1)}\right) + \mathsf{KL}\left(\pi^{(l_1+1)} \,\|\, \overline{\pi}^{(l_1-\gamma+1)}\right) \right] \\ &\quad + 2\eta \sum_{l_1=0}^{\gamma-1} (1-\eta\tau)^{\gamma-1-l_1} (l_1+1) d_{\mathsf{max}} \|A\|_\infty \\ &\le (1-\eta\tau)^\gamma \mathsf{KL}\left(\pi_\tau^\star \,\|\, \pi^{(0)}\right) - \sum_{l_1=0}^{\gamma-1} (1-\eta\tau)^{\gamma-1-l_1} \left[(1-\eta\tau)\mathsf{KL}\left(\overline{\pi}^{(l_1-\gamma+1)} \,\|\, \pi^{(l_1)}\right) + \mathsf{KL}\left(\pi^{(l_1+1)} \,\|\, \overline{\pi}^{(l_1-\gamma+1)}\right) \right] \\ &\quad + 2\eta\gamma^2 d_{\mathsf{max}} \|A\|_\infty.\end{aligned}$$

E.3.6 Proof of Lemma 40

Taking logarithm on the both sides of (E.37) and (E.38), we get

$$\eta\big(\log\widetilde{\pi}_i^{(t)} - \log\pi_i^{(t-1)}\big) \overset{\mathbf{1}}{=} \widetilde{\eta}_i^{(t)}\big(\log\pi_i^{(t)} - \log\pi_i^{(t-1)}\big),$$

or equivalently

$$\log\pi_i^{(t)} \overset{\mathbf{1}}{=} \frac{\eta}{\widetilde{\eta}_i^{(t)}}\log\widetilde{\pi}_i^{(t)} + \Big(1 - \frac{\eta}{\widetilde{\eta}_i^{(t)}}\Big)\log\pi_i^{(t-1)}.$$

Taking inner product with $\pi_{i,\tau}^\star - \pi_i^{(t)}$,

$$\bigg\langle \log\pi_i^{(t)} - \frac{\eta}{\widetilde{\eta}_i^{(t)}}\log\widetilde{\pi}_i^{(t)} - \Big(1 - \frac{\eta}{\widetilde{\eta}_i^{(t)}}\Big)\log\pi_i^{(t-1)},\ \pi_{i,\tau}^\star - \pi_i^{(t)} \bigg\rangle = 0.$$

By definition of KL divergence, we have

$$\begin{aligned}
&\bigg\langle \log\pi_i^{(t)} - \frac{\eta}{\widetilde{\eta}_i^{(t)}}\log\widetilde{\pi}_i^{(t)} - \Big(1 - \frac{\eta}{\widetilde{\eta}_i^{(t)}}\Big)\log\pi_i^{(t-1)},\ \pi_{i,\tau}^\star \bigg\rangle \\
&= \bigg\langle (\log\pi_i^{(t)} - \log\pi_{i,\tau}^\star) - \frac{\eta}{\widetilde{\eta}_i^{(t)}}(\log\widetilde{\pi}_i^{(t)} - \log\pi_{i,\tau}^\star) - \Big(1 - \frac{\eta}{\widetilde{\eta}_i^{(t)}}\Big)(\log\pi_i^{(t-1)} - \log\pi_{i,\tau}^\star),\ \pi_{i,\tau}^\star \bigg\rangle \\
&= -\mathsf{KL}\left(\pi_{i,\tau}^\star \,\|\, \pi_i^{(t)}\right) + \Big(1 - \frac{\eta}{\widetilde{\eta}_i^{(t)}}\Big)\mathsf{KL}\left(\pi_{i,\tau}^\star \,\|\, \pi_i^{(t-1)}\right) + \frac{\eta}{\widetilde{\eta}_i^{(t)}}\mathsf{KL}\left(\pi_{i,\tau}^\star \,\|\, \widetilde{\pi}_i^{(t)}\right),
\end{aligned}$$

and

$$\begin{aligned}
&\bigg\langle \log\pi_i^{(t)} - \frac{\eta}{\widetilde{\eta}_i^{(t)}}\log\widetilde{\pi}_i^{(t)} - \Big(1 - \frac{\eta}{\widetilde{\eta}_i^{(t)}}\Big)\log\pi_i^{(t-1)},\ \pi_i^{(t)} \bigg\rangle \\
&= \frac{\eta}{\widetilde{\eta}_i^{(t)}}\mathsf{KL}\left(\pi_i^{(t)} \,\|\, \widetilde{\pi}_i^{(t)}\right) + \Big(1 - \frac{\eta}{\widetilde{\eta}_i^{(t)}}\Big)\mathsf{KL}\left(\pi_i^{(t)} \,\|\, \pi_i^{(t-1)}\right).
\end{aligned}$$

Taken together, we get

$$\begin{aligned}
&\mathsf{KL}\left(\pi_{i,\tau}^\star \,\|\, \pi_i^{(t)}\right) + \frac{\eta}{\widetilde{\eta}_i^{(t)}}\mathsf{KL}\left(\pi_i^{(t)} \,\|\, \widetilde{\pi}_i^{(t)}\right) + \Big(1 - \frac{\eta}{\widetilde{\eta}_i^{(t)}}\Big)\mathsf{KL}\left(\pi_i^{(t)} \,\|\, \pi_i^{(t-1)}\right) \\
&= \Big(1 - \frac{\eta}{\widetilde{\eta}_i^{(t)}}\Big)\mathsf{KL}\left(\pi_{i,\tau}^\star \,\|\, \pi_i^{(t-1)}\right) + \frac{\eta}{\widetilde{\eta}_i^{(t)}}\mathsf{KL}\left(\pi_{i,\tau}^\star \,\|\, \widetilde{\pi}_i^{(t)}\right).
\end{aligned} \tag{E.61}$$

On the other hand, taking logarithm of (E.38) and making inner product with $\overline{\pi}_i^{(\kappa_i^{(t)})} - \pi_{i,\tau}^\star$ gives

$$\begin{aligned}
&\Big\langle \log\widetilde{\pi}_i^{(t)} - (1 - \widetilde{\eta}_i^{(t)}\tau)\log\pi_i^{(t-1)} - \widetilde{\eta}_i^{(t)}\tau\log\pi_{i,\tau}^\star,\ \overline{\pi}_i^{(\kappa_i^{(t)})} - \pi_{i,\tau}^\star \Big\rangle \\
&= \widetilde{\eta}_i^{(t)}(\overline{\pi}_i^{(\kappa_i^{(t)})} - \pi_{i,\tau}^\star)^\top A_i(\overline{\pi}^{(\kappa_i^{(t)})} - \pi_\tau^\star).
\end{aligned}$$

Following a similar discussion in (E.3) gives

$$\begin{aligned}
\mathsf{KL}\left(\pi_{i,\tau}^\star \,\|\, \widetilde{\pi}_i^{(t)}\right) &= (1 - \widetilde{\eta}_i^{(t)}\tau)\mathsf{KL}\left(\pi_{i,\tau}^\star \,\|\, \pi_i^{(t-1)}\right) - (1 - \widetilde{\eta}_i^{(t)}\tau)\mathsf{KL}\Big(\overline{\pi}_i^{(\kappa_i^{(t)})} \,\|\, \pi_i^{(t-1)}\Big) \\
&\quad - \widetilde{\eta}_i^{(t)}\tau\mathsf{KL}\Big(\overline{\pi}_i^{(\kappa_i^{(t)})} \,\|\, \pi_{i,\tau}^\star\Big) - \mathsf{KL}\Big(\widetilde{\pi}_i^{(t)} \,\|\, \overline{\pi}_i^{(\kappa_i^{(t)})}\Big) \\
&\quad + \Big\langle \log\overline{\pi}_i^{(\kappa_i^{(t)})} - \log\widetilde{\pi}_i^{(t)},\ \overline{\pi}_i^{(\kappa_i^{(t)})} - \widetilde{\pi}_i^{(t)} \Big\rangle - \widetilde{\eta}_i^{(t)}(\overline{\pi}_i^{(\kappa_i^{(t)})} - \pi_{i,\tau}^\star)^\top A_i(\overline{\pi}^{(\kappa_i^{(t)})} - \pi_\tau^\star).
\end{aligned} \tag{E.62}$$

Plugging the above equation into (E.61),

$$
\begin{aligned}
&\mathsf{KL}\left(\pi^\star_{i,\tau} \,\|\, \pi_i^{(t)}\right) + \frac{\eta}{\widetilde{\eta}_i^{(t)}}\mathsf{KL}\left(\pi_i^{(t)} \,\|\, \widetilde{\pi}_i^{(t)}\right) + \left(1 - \frac{\eta}{\widetilde{\eta}_i^{(t)}}\right)\mathsf{KL}\left(\pi_i^{(t)} \,\|\, \pi_i^{(t-1)}\right) \\
&= (1-\eta\tau)\mathsf{KL}\left(\pi^\star_{i,\tau} \,\|\, \pi_i^{(t-1)}\right) - \eta(\overline{\pi}_i^{(\kappa_i^{(t)})} - \pi^\star_{i,\tau})^\top A_i(\overline{\pi}^{(\kappa_i^{(t)})} - \pi^\star_\tau) \\
&\quad - \frac{\eta}{\widetilde{\eta}_i^{(t)}}\left[(1-\widetilde{\eta}_i^{(t)}\tau)\mathsf{KL}\left(\overline{\pi}_i^{(\kappa_i^{(t)})} \,\|\, \pi_i^{(t-1)}\right) + \widetilde{\eta}_i^{(t)}\tau\mathsf{KL}\left(\overline{\pi}_i^{(\kappa_i^{(t)})} \,\|\, \pi^\star_{i,\tau}\right) + \mathsf{KL}\left(\widetilde{\pi}_i^{(t)} \,\|\, \overline{\pi}_i^{(\kappa_i^{(t)})}\right)\right] \\
&\quad + \frac{\eta}{\widetilde{\eta}_i^{(t)}}\left\langle \log\overline{\pi}_i^{(\kappa_i^{(t)})} - \log\widetilde{\pi}_i^{(t)}, \overline{\pi}_i^{(\kappa_i^{(t)})} - \widetilde{\pi}_i^{(t)}\right\rangle.
\end{aligned}
$$

Rearranging the terms finishes the proof.

E.3.7 Proof of Lemma 41

For notational convenience, we set

$$
\begin{aligned}
\phi_i^{(t)} &= \left(1 - \frac{\eta}{\widetilde{\eta}_i^{(t)}}\right)\left\|\pi_i^{(t)} - \pi_i^{(t-1)}\right\|_1 \\
&\quad + \frac{\eta}{\widetilde{\eta}_i^{(t)}}\left(\left\|\overline{\pi}_i^{(\kappa_i^{(t)})} - \pi_i^{(t-1)}\right\|_1 + \left\|\widetilde{\pi}_i^{(t)} - \overline{\pi}_i^{(\kappa_i^{(t)})}\right\|_1 + \left\|\pi_i^{(t)} - \widetilde{\pi}_i^{(t)}\right\|_1\right)
\end{aligned}
$$

for all $i \in V, t \geq 0$. By triangular inequality, we have $\phi_i^{(t)} \geq \left\|\pi_i^{(t)} - \pi_i^{(t-1)}\right\|_1$. In addition, we denote by $t_1 \wedge t_2 := \min\{t_1, t_2\}$ and $t_1 \vee t_2 := \max\{t_1, t_2\}$. For $0 < t_1 < t_2$, it holds that

$$
\begin{aligned}
&\left\|\overline{\pi}_j^{(\kappa_i^{(t_1)})} - \overline{\pi}_j^{(\kappa_i^{(t_2)})}\right\|_1 \\
&\leq \left\|\pi_j^{(\nu_j(\kappa_i^{(t_1)}))} - \pi_j^{(\nu_j(\kappa_i^{(t_2)}))}\right\|_1 + \left\|\overline{\pi}_j^{(\kappa_i^{(t_1)})} - \pi_j^{(\nu_j(\kappa_i^{(t_1)}))}\right\|_1 + \left\|\overline{\pi}_j^{(\kappa_i^{(t_2)})} - \pi_j^{(\nu_j(\kappa_i^{(t_2)}))}\right\|_1 \\
&\leq \sum_{l=(\nu_j(\kappa_i^{(t_1)})+1)\wedge(\nu_j(\kappa_i^{(t_2)})+1)}^{\nu_j(\kappa_i^{(t_1)})\vee\nu_j(\kappa_i^{(t_2)})} \left\|\pi_j^{(l)} - \pi_j^{(l-1)}\right\|_1 + \left\|\overline{\pi}_j^{(\kappa_i^{(t_1)})} - \widetilde{\pi}_j^{(\nu_j(\kappa_i^{(t_1)}))}\right\|_1 + \left\|\widetilde{\pi}_j^{(\nu_j(\kappa_i^{(t_1)}))} - \pi_j^{(\nu_j(\kappa_i^{(t_1)}))}\right\|_1 \\
&\quad + \left\|\overline{\pi}_j^{(\kappa_i^{(t_2)})} - \widetilde{\pi}_j^{(\nu_j(\kappa_i^{(t_2)}))}\right\|_1 + \left\|\widetilde{\pi}_j^{(\nu_j(\kappa_i^{(t_2)}))} - \pi_j^{(\nu_j(\kappa_i^{(t_2)}))}\right\|_1 \\
&\leq \sum_{l=(\nu_j(\kappa_i^{(t_1)})+1)\wedge(\nu_j(\kappa_i^{(t_2)})+1)}^{\nu_j(\kappa_i^{(t_1)})\vee\nu_j(\kappa_i^{(t_2)})} \phi_j^{(l)} + \frac{\widetilde{\eta}_j^{(\nu_j(\kappa_i^{(t_1)}))}}{\eta}\phi_j^{(\nu_j(\kappa_i^{(t_1)}))} + \frac{\widetilde{\eta}_j^{(\nu_j(\kappa_i^{(t_2)}))}}{\eta}\phi_j^{(\nu_j(\kappa_i^{(t_2)}))} \\
&\leq \sum_{l=(\nu_j(\kappa_i^{(t_1)})+1)\wedge(\nu_j(\kappa_i^{(t_2)})+1)}^{\nu_j(\kappa_i^{(t_1)})\vee\nu_j(\kappa_i^{(t_2)})} \phi_j^{(l)} + (\gamma+1)\phi_j^{(\nu_j(\kappa_i^{(t_1)}))} + (\gamma+1)\phi_j^{(\nu_j(\kappa_i^{(t_2)}))}. \qquad \text{(E.63)}
\end{aligned}
$$

Therefore, we have

$$\sum_{k=\kappa_i^{(t)}}^{t} \left\|\overline{\pi}_j^{(\kappa_i^{(k)})} - \overline{\pi}_j^{(\kappa_i^{(\kappa_i^{(t-1)})})}\right\|_1$$
$$\leq \sum_{k=\kappa_i^{(t)}}^{t} \left\{ \sum_{l=(\nu_j(\kappa_i^{(\kappa_i^{(t-1)})})+1)\wedge(\nu_j(\kappa_i^{(k)})+1)}^{\nu_j(\kappa_i^{(\kappa_i^{(t-1)})})\vee\nu_j(\kappa_i^{(k)})} \phi_j^{(l)} + (\gamma+1)\phi_j^{(\nu_j(\kappa_i^{(\kappa_i^{(t-1)})}))} + (\gamma+1)\phi_j^{(\nu_j(\kappa_i^{(k)}))} \right\}. \quad \text{(E.64)}$$

Since $0 \vee (t-\gamma) \leq \kappa_i^{(t)} \leq t \leq \nu_i(t) \leq t+\gamma$ for all $i \in V$, $t \geq 0$, the first term can be bounded by

$$\sum_{l=(\nu_j(\kappa_i^{(\kappa_i^{(t-1)})})+1)\wedge(\nu_j(\kappa_i^{(k)})+1)}^{\nu_j(\kappa_i^{(\kappa_i^{(t-1)})})\vee\nu_j(\kappa_i^{(k)})} \phi_j^{(l)} \leq \sum_{l=(t-2\gamma)\wedge(k-\gamma+1)}^{(t+\gamma-1)\vee(k+\gamma)} \phi_j^{(l)} \leq \sum_{l=t-2\gamma}^{t+\gamma} \phi_j^{(l)}.$$

In addition, the mapping $k \mapsto \nu_j(\kappa_i^{(k)})$ is injective when $k \geq \gamma$ (cf. Assumption 5 and 6). It follows that

$$\sum_{k=\kappa_i^{(t)}}^{t} \phi_j^{(\nu_j(\kappa_i^{(k)}))} \leq \sum_{l=\kappa_i^{(t)}-\gamma}^{t+\gamma} \phi_j^{(l)} \leq \sum_{l=t-2\gamma}^{t+\gamma} \phi_j^{(l)}$$

Plugging the above inequalities into (E.64) yields

$$\sum_{k=\kappa_i^{(t)}}^{t} \left\|\overline{\pi}_j^{(\kappa_i^{(k)})} - \overline{\pi}_j^{(\kappa_i^{(\kappa_i^{(t-1)})})}\right\|_1$$
$$\leq (t+1-\kappa_i^{(t)}) \sum_{l=t-2\gamma}^{t+\gamma} \phi_j^{(l)} + (t+1-\kappa_i^{(t)})(\gamma+1)\phi_j^{(\nu_j(\kappa_i^{(\kappa_i^{(t-1)})}))} + (\gamma+1) \sum_{l=t-2\gamma}^{t+\gamma} \phi_j^{(l)}$$
$$\leq 2(\gamma+1) \sum_{l=t-2\gamma}^{t+\gamma} \phi_j^{(l)} + (\gamma+1)^2 \phi_j^{(\nu_j(\kappa_i^{(\kappa_i^{(t-1)})}))}.$$

Finally, we control the term $\phi_i^{(t)}$ with $\psi_i^{(t)}$ as:

$$\begin{aligned}
(\phi_i^{(t)})^2 &= \Bigg(\Big(1-\frac{\eta}{\widetilde{\eta}_i^{(t)}}\Big)^{1/2}\cdot\Big(1-\frac{\eta}{\widetilde{\eta}_i^{(t)}}\Big)^{1/2}\big\|\pi_i^{(t)}-\pi_i^{(t-1)}\big\|_1 \\
&\quad+\Big(\frac{\eta}{\widetilde{\eta}_i^{(t)}}(1-\widetilde{\eta}_i^{(t)}\tau)^{-1}\Big)^{1/2}\cdot\Big(\frac{\eta}{\widetilde{\eta}_i^{(t)}}(1-\widetilde{\eta}_i^{(t)}\tau)\Big)^{1/2}\big\|\overline{\pi}_i^{(\kappa_i^{(t)})}-\pi_i^{(t-1)}\big\|_1 \\
&\quad+\Big(\frac{\eta}{\widetilde{\eta}_i^{(t)}}\Big)^{1/2}\cdot\Big(\frac{\eta}{\widetilde{\eta}_i^{(t)}}\Big)^{1/2}\Big(\big\|\widetilde{\pi}_i^{(t)}-\overline{\pi}_i^{(\kappa_i^{(t)})}\big\|_1+\big\|\pi_i^{(t)}-\widetilde{\pi}_i^{(t)}\big\|_1\Big)\Bigg)^2 \\
&\overset{\text{(i)}}{\le}\Big(1-\frac{\eta}{\widetilde{\eta}_i^{(t)}}+\frac{\eta}{\widetilde{\eta}_i^{(t)}}\big(2+(1-\widetilde{\eta}_i^{(t)}\tau)^{-1}\big)\Big)\Bigg[\Big(1-\frac{\eta}{\widetilde{\eta}_i^{(t)}}\Big)\big\|\pi_i^{(t)}-\pi_i^{(t-1)}\big\|_1^2 \\
&\quad+\frac{\eta}{\widetilde{\eta}_i^{(t)}}\Big((1-\widetilde{\eta}_i^{(t)}\tau)\big\|\overline{\pi}_i^{(\kappa_i^{(t)})}-\pi_i^{(t-1)}\big\|_1^2+\big\|\widetilde{\pi}_i^{(t)}-\overline{\pi}_i^{(\kappa_i^{(t)})}\big\|_1^2+\big\|\pi_i^{(t)}-\widetilde{\pi}_i^{(t)}\big\|_1^2\Big)\Bigg] \\
&\overset{\text{(ii)}}{\le} 2\,\big(2+(1-\widetilde{\eta}_i^{(t)}\tau)^{-1}\big)\Bigg[\Big(1-\frac{\eta}{\widetilde{\eta}_i^{(t)}}\Big)\mathsf{KL}\left(\pi_i^{(t)}\,\|\,\pi_i^{(t-1)}\right) \\
&\quad+\frac{\eta}{\widetilde{\eta}_i^{(t)}}\Big((1-\widetilde{\eta}_i^{(t)}\tau)\mathsf{KL}\Big(\overline{\pi}_i^{(\kappa_i^{(t)})}\,\|\,\pi_i^{(t-1)}\Big)+\mathsf{KL}\Big(\widetilde{\pi}_i^{(t)}\,\|\,\overline{\pi}_i^{(\kappa_i^{(t)})}\Big)+\mathsf{KL}\left(\pi_i^{(t)}\,\|\,\widetilde{\pi}_i^{(t)}\right)\Big)\Bigg] \\
&\overset{\text{(iii)}}{\le} 8\psi_i^{(t)},
\end{aligned}\tag{E.65}$$

where (i) applies Cauchy-Schwarz inequality, (ii) invokes Pinsker's inequality and (iii) is due to $\widetilde{\eta}_i^{(t)}\tau\le(\gamma+1)\eta\tau\le 1/2$. Combining the above two inequalities finishes the proof.

E.3.8 Proof of Lemma 42

We start with verifying the claim (E.46). Recall that

$$\begin{aligned}
\psi_i^{(t)} &:= \Big(1-\frac{\eta}{\widetilde{\eta}_i^{(t)}}\Big)\mathsf{KL}\left(\pi_i^{(t)}\,\|\,\pi_i^{(t-1)}\right) \\
&\quad+\frac{\eta}{\widetilde{\eta}_i^{(t)}}\left[(1-\widetilde{\eta}_i^{(t)}\tau)\mathsf{KL}\Big(\overline{\pi}_i^{(\kappa_i^{(t)})}\,\|\,\pi_i^{(t-1)}\Big)+\mathsf{KL}\Big(\widetilde{\pi}_i^{(t)}\,\|\,\overline{\pi}_i^{(\kappa_i^{(t)})}\Big)+\mathsf{KL}\left(\pi_i^{(t)}\,\|\,\widetilde{\pi}_i^{(t)}\right)\right].
\end{aligned}$$

We introduce the following standard Lemma (c.f., (A.34)), which allows us to bound control $\mathsf{KL}\left(\pi_i\,\|\,\pi_i'\right)$ properly:

Lemma 44. *Given $\pi_i,\pi_i'\in\Delta(S_i)$ and $w\in\mathbb{R}^{|S_i|}$ with $\log\pi_i\overset{\mathbf{1}}{=}\log\pi_i'+w$, we have*

$$\mathsf{KL}\left(\pi_i\,\|\,\pi_i'\right)\le\big\|\log\pi_i-\log\pi_i'\big\|_\infty\le 2\big\|w\big\|_\infty.$$

Therefore, it suffices to figure out the terms $\log\pi_i^{(t)}-\log\pi_i^{(t-1)}$, $\log\pi_i^{(t)}-\log\widetilde{\pi}_i^{(t)}$, $\log\widetilde{\pi}_i^{(t)}-\log\overline{\pi}_i^{(\kappa_i^{(t)})}$ and $\log\overline{\pi}_i^{(\kappa_i^{(t)})}-\log\pi_i^{(t-1)}$.

- **Bounding $\mathsf{KL}\left(\pi_i^{(t)}\,\|\,\pi_i^{(t-1)}\right)$ and $\mathsf{KL}\left(\pi_i^{(t)}\,\|\,\widetilde{\pi}_i^{(t)}\right)$.** The following equations follow directly from (E.37) and (E.38):

$$\begin{cases}\log\pi_i^{(t)}-\log\pi_i^{(t-1)}\overset{\mathbf{1}}{=}\eta([A_i\overline{\pi}^{(\kappa_i^{(t)})}]_k-\tau\log\pi_i^{(t-1)}) \\ \log\pi_i^{(t)}-\log\widetilde{\pi}_i^{(t)}\overset{\mathbf{1}}{=}(\eta-\widetilde{\eta}_i^{(t)})([A_i\overline{\pi}^{(\kappa_i^{(t)})}]_k-\tau\log\pi_i^{(t-1)})\end{cases}.\tag{E.66}$$

In addition, we have the following bound w.r.t. the order of $\left\|\log \pi_i^{(t-1)}\right\|_\infty$, which we shall establish momentarily.

$$\left\|\tau \log \pi_i^{(t-1)}\right\|_\infty \le \tau \log |S_i| + 2d_{\mathsf{max}} \|A\|_\infty . \tag{E.67}$$

This taken together with Lemma 44 yields

$$\begin{cases} \mathsf{KL}\left(\pi_i^{(t)} \,\|\, \pi_i^{(t-1)}\right) \le \eta(3d_{\mathsf{max}} \|A\|_\infty + \tau \log |S_i|) \\ \mathsf{KL}\left(\pi_i^{(t)} \,\|\, \widetilde{\pi}_i^{(t)}\right) \le (\widetilde{\eta}_i^{(t)} - \eta)(3d_{\mathsf{max}} \|A\|_\infty + \tau \log |S_i|) \end{cases}. \tag{E.68}$$

- **Bounding $\mathsf{KL}\left(\widetilde{\pi}_i^{(t)} \,\|\, \overline{\pi}_i^{(\kappa_i^{(t)})}\right)$.** When $\kappa_i^{(t)} \ge 1$, we recall from (E.40) that:

$$\begin{aligned} \log \overline{\pi}_i^{(\kappa_i^{(t)})} - \log \widetilde{\pi}_i^{(t)} \overset{\mathbf{1}}{=} \widetilde{\eta}_i^{(t)} \Bigg(& A_i(\overline{\pi}^{(\kappa_i^{(\kappa_i^{(t)}-1)})} - \overline{\pi}^{(\kappa_i^{(t)})}) \\ & + \sum_{l=\kappa_i^{(t)}}^{t-1} (1-\widetilde{\eta}_i^{(t)}\tau)(1-\eta\tau)^{t-1-l} A_i(\overline{\pi}^{(\kappa_i^{(\kappa_i^{(t)}-1)})} - \overline{\pi}^{(\kappa_i^{(l)})}) \Bigg), \end{aligned} \tag{E.69}$$

which leads to a crude bound

$$\mathsf{KL}\left(\widetilde{\pi}_i^{(t)} \,\|\, \overline{\pi}_i^{(\kappa_i^{(t)})}\right) \le \widetilde{\eta}_i^{(t)} d_{\mathsf{max}} \|A\|_\infty (t - \kappa_i^{(t)} + 1) \le \widetilde{\eta}_i^{(t)} d_{\mathsf{max}} \|A\|_\infty (\gamma + 1).$$

When $\kappa_i^{(t)} = 0$, we have

$$\begin{aligned} \log \overline{\pi}_i^{(\kappa_i^{(t)})} - \log \widetilde{\pi}_i^{(t)} & \overset{\mathbf{1}}{=} -\log \widetilde{\pi}_i^{(t)} \\ & \overset{\mathbf{1}}{=} -(1-\widetilde{\eta}_i^{(t)}\tau)(1-\eta\tau)^{t-1} \log \pi^{(0)} \\ & \quad - \widetilde{\eta}_i^{(t)} \left(A_i \overline{\pi}^{(\kappa_i^{(t)})} + \sum_{l=\kappa_i^{(t)}+1}^{t-1} (1-\widetilde{\eta}_i^{(t)}\tau)(1-\eta\tau)^{t-1-l} A_i \overline{\pi}^{(\kappa_i^{(l)})} \right) \\ & \overset{\mathbf{1}}{=} -\widetilde{\eta}_i^{(t)} \left(A_i \overline{\pi}^{(\kappa_i^{(t)})} + \sum_{l=\kappa_i^{(t)}+1}^{t-1} (1-\widetilde{\eta}_i^{(t)}\tau)(1-\eta\tau)^{t-1-l} A_i \overline{\pi}^{(\kappa_i^{(l)})} \right), \end{aligned} \tag{E.70}$$

which yields

$$\mathsf{KL}\left(\widetilde{\pi}_i^{(t)} \,\|\, \overline{\pi}_i^{(\kappa_i^{(t)})}\right) \le \widetilde{\eta}_i^{(t)} d_{\mathsf{max}} \|A\|_\infty t \le \widetilde{\eta}_i^{(t)} d_{\mathsf{max}} \|A\|_\infty (\gamma + 1).$$

- **Bounding $\mathsf{KL}\left(\overline{\pi}_i^{(\kappa_i^{(t)})} \,\|\, \pi_i^{(t-1)}\right)$.** Note that

$$\log \overline{\pi}_i^{(\kappa_i^{(t)})} - \log \pi_i^{(t-1)} = (\log \overline{\pi}_i^{(\kappa_i^{(t)})} - \log \widetilde{\pi}_i^{(t)}) + (\log \widetilde{\pi}_i^{(t)} - \log \pi_i^{(t)}) + (\log \pi_i^{(t)} - \log \pi_i^{(t-1)}).$$

This yields, by equations (E.66), (E.69), (E.70) and associated bounds,

$$\mathsf{KL}\left(\overline{\pi}_i^{(\kappa_i^{(t)})} \,\|\, \pi_i^{(t-1)}\right) \le \widetilde{\eta}_i^{(t)} d_{\mathsf{max}} \|A\|_\infty (\gamma + 1) + \widetilde{\eta}_i^{(t)} (3d_{\mathsf{max}} \|A\|_\infty + \tau \log |S_i|).$$

Putting all pieces together, we conclude that

$$\begin{aligned}\psi_i^{(t)} &= \Big(1-\frac{\eta}{\widetilde{\eta}_i^{(t)}}\Big)\mathsf{KL}\left(\pi_i^{(t)} \,\|\, \pi_i^{(t-1)}\right) \\ &\quad + \frac{\eta}{\widetilde{\eta}_i^{(t)}}\left[(1-\widetilde{\eta}_i^{(t)}\tau)\mathsf{KL}\Big(\overline{\pi}_i^{(\kappa_i^{(t)})} \,\|\, \pi_i^{(t-1)}\Big) + \mathsf{KL}\Big(\widetilde{\pi}_i^{(t)} \,\|\, \overline{\pi}_i^{(\kappa_i^{(t)})}\Big) + \mathsf{KL}\left(\pi_i^{(t)} \,\|\, \widetilde{\pi}_i^{(t)}\right)\right] \\ &\le 3\eta(3d_{\mathsf{max}}\left\|A\right\|_\infty + \tau\log|S_i|) + 2\eta d_{\mathsf{max}}\left\|A\right\|_\infty(\gamma+1) \\ &= \eta(d_{\mathsf{max}}\left\|A\right\|_\infty(2\gamma+11) + 3\tau\log|S_i|).\end{aligned}$$

It remains to prove the claim (E.47):

$$\begin{aligned}\mathsf{KL}\left(\pi_{i,\tau}^\star \,\|\, \pi_i^{(2\gamma)}\right) &= \mathsf{KL}\left(\pi_{i,\tau}^\star \,\|\, \pi_i^{(0)}\right) + \left\langle \pi_{i,\tau}^\star, \log\pi_i^{(0)} - \log\pi_i^{(2\gamma)}\right\rangle \\ &\le \mathsf{KL}\left(\pi_{i,\tau}^\star \,\|\, \pi_i^{(0)}\right) + \left\|\log\pi_i^{(0)} - \log\pi_i^{(2\gamma)}\right\|_\infty \\ &\le \mathsf{KL}\left(\pi_{i,\tau}^\star \,\|\, \pi_i^{(0)}\right) + 2\left\|\eta\sum_{l=1}^{2\gamma}(1-\eta\tau)^{2\gamma-l}A_i\overline{\pi}^{(\kappa_i^{(l)})}\right\|_\infty \\ &\le \mathsf{KL}\left(\pi_{i,\tau}^\star \,\|\, \pi_i^{(0)}\right) + 4\eta d_{\mathsf{max}}\left\|A\right\|_\infty\gamma,\end{aligned}$$

where the third step results from $\log\pi_i^{(2\gamma)} \overset{\mathbf{1}}{=} (1-\eta\tau)^{2\gamma}\log\pi_i^{(0)} + \eta\sum_{l=1}^{2\gamma}(1-\eta\tau)^{2\gamma-l}A_i\overline{\pi}^{(\kappa_i^{(l)})}$ and Lemma 44.

Proof of the claim (E.67). First, we prove by induction that for any $k, l \in S_i$,

$$\log\pi_i^{(t)}(k) - \log\pi_i^{(t)}(l) \le \frac{2d_{\mathsf{max}}\left\|A\right\|_\infty}{\tau}, \qquad \forall t \ge 0. \tag{E.71}$$

Note that the claim trivially holds for $t=0$ with the uniform initialization $\pi_i^{(0)} = \frac{1}{|S_i|}\mathbf{1}$, $\forall i \in V$. Assume that (E.71) holds for all $t' \le t-1$. Note that $\log\pi_i^{(t)} \overset{\mathbf{1}}{=} (1-\eta\tau)\log\pi_i^{(t-1)} + \eta A_i\overline{\pi}^{(\kappa_i^{(t)})}$, we have

$$\begin{aligned}\log\pi_i^{(t)}(k) - \log\pi_i^{(t)}(l) &= (1-\eta\tau)\left(\log\pi_i^{(t-1)}(k) - \log\pi_i^{(t-1)}(l)\right) + \eta\left([A_i\overline{\pi}^{(\kappa_i^{(t)})}]_k - [A_i\overline{\pi}^{(\kappa_i^{(t)})}]_l\right) \\ &\le (1-\eta\tau)\frac{2d_{\mathsf{max}}\left\|A\right\|_\infty}{\tau} + 2\eta d_{\mathsf{max}}\left\|A\right\|_\infty \\ &= \frac{2d_{\mathsf{max}}\left\|A\right\|_\infty}{\tau},\end{aligned}$$

where the second line follows from the induction hypothesis (E.71). This completes the induction at the t-th iteration. It follows that for all $i \in V$ and $t \ge 0$,

$$\log\pi_i^{(t)}(l) \ge \log\left(\max_{k\in S_i}\pi_i^{(t)}(k)\right) - \frac{2d_{\mathsf{max}}\left\|A\right\|_\infty}{\tau} \ge -\log|S_i| - \frac{2d_{\mathsf{max}}\left\|A\right\|_\infty}{\tau}. \tag{E.72}$$

Appendix F

Proofs for Chapter 7

F.1 Proof of Theorem 12

Before proceeding to the main proof, we first a useful lemma connecting the marginalized utility with the policy, as given below

Lemma 45. *Given any $\pi, \pi' \in \Delta(\mathcal{A})^N$, the difference in the marginalized utility (cf. (7.2)) can be bounded by*

$$\left\| r_i^{\pi} - r_i^{\pi'} \right\|_{\infty} \leq \sqrt{J(\pi, \pi')},$$

where $J(\pi, \pi') = \mathsf{KL}\left(\pi \,\|\, \pi'\right) + \mathsf{KL}\left(\pi' \,\|\, \pi\right)$ is the Jeffrey divergence.

Proof. See Appendix F.2. □

F.1.1 Step 1: quantify the policy improvement

We start by the following key lemma that gives a lower bound of the improvement in terms of the regularized potential function $\Phi_{\tau}^{(t)}$.

Lemma 46. *The independent NPG update (7.5) guarantees that*

$$\Phi_{\tau}^{(t+1)} - \Phi_{\tau}^{(t)} \geq \left(\frac{1}{\eta} - \min\{\sqrt{N}, 2\Phi_{\max}\} - \tau \right) J\left(\pi^{(t+1)}, \pi^{(t)}\right).$$

Proof. See Appendix F.3. □

Lemma 46 ensures the monotonic improvement of the regularized potential function Φ_{τ} when η is not too large. Specifically, setting $\eta \leq 1/(2(\min\{\sqrt{N}, 2\Phi_{\max}\} + \tau))$, we have

$$\Phi_{\tau}^{(t+1)} - \Phi_{\tau}^{(t)} \geq \frac{1}{2\eta} J\left(\pi^{(t+1)}, \pi^{(t)}\right),$$

which is guaranteed to be non-negative. Summing the above inequality over $t = 0, \cdots, T-1$ gives

$$\sum_{t=0}^{T-1} J\left(\pi^{(t+1)}, \pi^{(t)}\right) \leq 2\eta(\Phi_{\tau}^{(T)} - \Phi_{\tau}^{(0)}), \tag{F.1}$$

which controls the change of $\pi^{(t)}$ over time t via the size of the regularized potential function.

F.1.2 Step 2: introduce the auxiliary sequence

Motivated by Cen et al. [2021, 2022b], we introduce an auxiliary sequence $\{\xi_i^{(t)} \in \mathbb{R}^{|\mathcal{A}|},\ i \in [N]\}$, constructed recursively by

$$\xi_i^{(0)}(a) = \left\|\exp(r_i^{(0)}/\tau)\right\|_1 \cdot \pi_i^{(0)}(a), \tag{F.2a}$$

$$\xi_i^{(t+1)}(a) = \xi_i^{(t)}(a)^{1-\eta\tau} \exp(\eta r_i^{(t)}(a)). \tag{F.2b}$$

Compared with the independent NPG update rule (7.5), it is clear that $\xi_i^{(t)} \propto \pi_i^{(t)}$ up to normalization. In addition, we have

$$\begin{aligned}\log \xi_i^{(t+1)} - r_i^{(t+1)}/\tau &= (1-\eta\tau)\log \xi_i^{(t)} + \eta r_i^{(t)} - r_i^{(t+1)}/\tau \\ &= (1-\eta\tau)\left(\log \xi_i^{(t)} - r_i^{(t)}/\tau\right) + \left(r_i^{(t)} - r_i^{(t+1)}\right)/\tau,\end{aligned}$$

which implies

$$\begin{aligned}\left\|\log \xi_i^{(t+1)} - r_i^{(t+1)}/\tau\right\|_\infty &\le (1-\eta\tau)\left\|\log \xi_i^{(t)} - r_i^{(t)}/\tau\right\|_\infty + \left\|r_i^{(t)} - r_i^{(t+1)}\right\|_\infty/\tau \\ &\le (1-\eta\tau)^{t+1}\left\|\log \xi_i^{(0)} - r_i^{(0)}/\tau\right\|_\infty + \tau^{-1}\sum_{s=0}^{t}(1-\eta\tau)^{t-s}\left\|r_i^{(s)} - r_i^{(s+1)}\right\|_\infty \\ &\le (1-\eta\tau)^{t+1}\left\|\log \xi_i^{(0)} - r_i^{(0)}/\tau\right\|_\infty + \tau^{-1}\sum_{s=0}^{t}(1-\eta\tau)^{t-s}\sqrt{J\left(\pi^{(s+1)}, \pi^{(s)}\right)},\end{aligned} \tag{F.3}$$

where the last line follows by applying Lemma 45 to the last term by setting $\pi = \pi^{(t)}$ and $\pi' = \pi^{(t+1)}$.

F.1.3 Step 3: bound the gap

Note that by the definition of the best-response policy in (7.6), the term of interest in $\texttt{QRE-gap}_\tau^{(t)}$ can be controlled as

$$\begin{aligned}u_{i,\tau}(\pi_i^{\star(t+1)}, \pi_{-i}^{(t+1)}) - u_{i,\tau}(\pi_i^{(t+1)}, \pi_{-i}^{(t+1)}) &= \left\langle \pi_i^{\star(t+1)} - \pi_i^{(t+1)}, r_i^{(t+1)}\right\rangle + \tau\mathcal{H}(\pi_i^{\star(t+1)}) - \tau\mathcal{H}(\pi_i^{(t+1)}) \\ &= \tau\mathsf{KL}\left(\pi_i^{(t+1)} \,\|\, \pi_i^{\star(t+1)}\right) \le \tau\left\|\log \pi_i^{(t+1)} - \log \pi_i^{\star(t+1)}\right\|_\infty \\ &\le 2\tau\left\|\log \xi_i^{(t+1)} - r_i^{(t+1)}/\tau\right\|_\infty,\end{aligned}$$

where the first line follows from the definition (7.2), the second step results from a direct consequence of (7.7):

$$\left\langle \pi_i^{\star(t+1)} - \pi_i^{(t+1)}, r_i^{(t+1)}\right\rangle = \left\langle \pi_i^{\star(t+1)} - \pi_i^{(t+1)}, \tau\log \pi_i^{\star(t+1)}\right\rangle$$

with a little algebra, and the last line follows from Lemma 15. Taking maximum over $i \in [N]$, in conjunction with (F.3), we end up with

$$\texttt{QRE-gap}_\tau^{(t+1)} \le 2\tau(1-\eta\tau)^{t+1}\left\|\log \pi^{(0)} - \log \pi^{\star(0)}\right\|_\infty + 2\sum_{s=0}^{t}(1-\eta\tau)^{t-s}\sqrt{J\left(\pi^{(s+1)}, \pi^{(s)}\right)}.$$

Summing the inequality over $t = 0, \ldots, T-1$ gives

$$\begin{aligned}
&\sum_{t=0}^{T-1} \texttt{QRE-gap}_\tau^{(t+1)} \\
&\leq 2\tau \sum_{t=0}^{T-1} (1-\eta\tau)^{t+1} \max_{i\in[N]} \left\| \log \pi_i^{(0)} - \log \pi_i^{\star(0)} \right\|_\infty + 2 \sum_{t=0}^{T-1} \sum_{s=0}^{t} (1-\eta\tau)^{t-s} \sqrt{J\left(\pi^{(s+1)}, \pi^{(s)}\right)} \\
&\leq \frac{2}{\eta\tau} \Big(\tau \left\| \log \pi^{(0)} - \log \pi^{\star(0)} \right\|_\infty + \sum_{s=0}^{T-1} \sqrt{J\left(\pi^{(s+1)}, \pi^{(s)}\right)} \Big).
\end{aligned}$$

The proof is thus completed by noticing

$$\sum_{s=0}^{T-1} \sqrt{J\left(\pi^{(s+1)}, \pi^{(s)}\right)} \leq \sqrt{T \sum_{s=0}^{T-1} J\left(\pi^{(s+1)}, \pi^{(s)}\right)} \leq \sqrt{2\eta T (\Phi_\tau^{(T)} - \Phi_\tau^{(0)})}.$$

Here, the second step results from Pinsker's inequality, and the last line follows from (F.1).

F.2 Proof of Lemma 45

Given any $\pi, \pi' \in \Delta(\mathcal{A})^N$, we have

$$\begin{aligned}
\left| r_i^{\pi}(a) - r_i^{\pi'}(a) \right| &= \left| \mathbb{E}_{a_{-i} \sim \pi_{-i}} \left[u_i(a,\, a_{-i}) \right] - \mathbb{E}_{a_{-i} \sim \pi'_{-i}} \left[u_i(a,\, a_{-i}) \right] \right| \\
&\overset{\text{(i)}}{\leq} 2 \left\| u_i \right\|_\infty d_{TV}\left(\pi_{-i},\, \pi'_{-i}\right) \\
&\leq \sqrt{2 d_{TV}\left(\pi_{-i},\, \pi'_{-i}\right)^2 + 2 d_{TV}\left(\pi'_{-i},\, \pi_{-i}\right)^2} \\
&\overset{\text{(ii)}}{\leq} \sqrt{\mathsf{KL}\left(\pi_{-i} \,\|\, \pi'_{-i}\right) + \mathsf{KL}\left(\pi'_{-i} \,\|\, \pi_{-i}\right)} \\
&\leq \sqrt{\mathsf{KL}\left(\pi \,\|\, \pi'\right) + \mathsf{KL}\left(\pi' \,\|\, \pi\right)} = \sqrt{J(\pi, \pi')},
\end{aligned} \tag{F.4}$$

where $d_{TV}(\cdot,\, \cdot)$ refers to total variation distance. Here, (i) follows from applying $\left| \int_\Omega h \mathrm{d}\mu - \int_\Omega h \mathrm{d}\nu \right| \leq 2 d_{TV}(\mu, \nu) \left\| h \right\|_\infty$ which holds for any probability measures μ, ν and bounded measurable function $h : \Omega \to \mathbb{R}$ (see e.g., [Driver, 2007, Corollary 13.4]), and (ii) results from Pinsker's inequality.

F.3 Proof of Lemma 46

The proof is composed of two parts, each establishing the following bounds

$$\Phi_\tau^{(t+1)} - \Phi_\tau^{(t)} \geq \left(\frac{1}{\eta} - \sqrt{N} - \tau \right) J\left(\pi^{(t+1)}, \pi^{(t)}\right), \tag{F.5a}$$

$$\Phi_\tau^{(t+1)} - \Phi_\tau^{(t)} \geq \left(\frac{1}{\eta} - 2\Phi_{\max} - \tau \right) J\left(\pi^{(t+1)}, \pi^{(t)}\right) \tag{F.5b}$$

respectively. Combining the two bounds then finishes the proof.

F.3.1 Proof of (F.5a)

We introduce

$$\widetilde{\pi}_{-i}^{(t)}(a_{-i}) = \prod_{j<i} \pi_j^{(t)}(a_j) \prod_{k>i} \pi_k^{(t+1)}(a_k) \in \Delta(\mathcal{A})^{N-1}$$

to denote the mixed strategy profile (except that of agent i) where the agents with index $j < i$ follow $\pi_j^{(t)}$ and the agents with index $k > i$ follow $\pi_k^{(t+1)}$ instead. Let $\widetilde{r}_i^{(t)}$ be the associated marginalized utility function, i.e.,

$$\begin{aligned} \widetilde{r}_i^{(t)}(a) &= \mathop{\mathbb{E}}_{a_i=a, a_{-i}\sim\widetilde{\pi}_{-i}^{(t)}} [u_i(\boldsymbol{a})] \\ &= \sum_{a_{-i}\in\mathcal{A}^{N-1}} u_i(a,\, a_{-i}) \prod_{j<i} \pi_j^{(t)}(a_j) \prod_{k>i} \pi_k^{(t+1)}(a_k). \end{aligned} \tag{F.6}$$

It follows from the above definition that we have

$$\Phi_\tau(\pi_i^{(t)}, \widetilde{\pi}_{-i}^{(t)}) = \Phi_\tau(\pi_1^{(t)}, \cdots, \pi_i^{(t)}, \pi_{i+1}^{(t+1)}, \cdots, \pi_N^{(t+1)}) = \Phi_\tau(\pi_{i+1}^{(t+1)}, \widetilde{\pi}_{-(i+1)}^{(t)}) \tag{F.7}$$

for $i \in [N-1]$.

We now decompose $\Phi_\tau^{(t+1)} - \Phi_\tau^{(t)}$ as follows:

$$\begin{aligned} \Phi_\tau^{(t+1)} - \Phi_\tau^{(t)} &= \Phi_\tau(\pi_1^{(t+1)}, \widetilde{\pi}_{-1}^{(t)}) - \Phi_\tau(\pi_N^{(t)}, \widetilde{\pi}_{-N}^{(t)}) \\ &= \Phi_\tau(\pi_1^{(t+1)}, \widetilde{\pi}_{-1}^{(t)}) - \Phi_\tau(\pi_1^{(t)}, \widetilde{\pi}_{-1}^{(t)}) + \Phi_\tau(\pi_1^{(t)}, \widetilde{\pi}_{-1}^{(t)}) - \Phi_\tau(\pi_N^{(t)}, \widetilde{\pi}_{-N}^{(t)}) \\ &\overset{\text{(i)}}{=} \left[\Phi_\tau(\pi_1^{(t+1)}, \widetilde{\pi}_{-1}^{(t)}) - \Phi_\tau(\pi_1^{(t)}, \widetilde{\pi}_{-1}^{(t)})\right] + \Phi_\tau(\pi_2^{(t+1)}, \widetilde{\pi}_{-2}^{(t)}) - \Phi_\tau(\pi_N^{(t)}, \widetilde{\pi}_{-N}^{(t)}) \\ &\overset{\text{(ii)}}{=} \sum_{i=1}^N \left[\Phi_\tau(\pi_i^{(t+1)}, \widetilde{\pi}_{-i}^{(t)}) - \Phi_\tau(\pi_i^{(t)}, \widetilde{\pi}_{-i}^{(t)})\right] \\ &= \sum_{i=1}^N \left[u_{i,\tau}(\pi_i^{(t+1)}, \widetilde{\pi}_{-i}^{(t)}) - u_{i,\tau}(\pi_i^{(t)}, \widetilde{\pi}_{-i}^{(t)})\right] \\ &= \sum_{i=1}^N \left[\left\langle \widetilde{r}_i^{(t)}, \pi_i^{(t+1)} - \pi_i^{(t)} \right\rangle + \tau \left(\mathcal{H}(\pi_i^{(t+1)}) - \mathcal{H}(\pi_i^{(t)})\right)\right], \end{aligned}$$

where (i) follows from (F.7), and (ii) follows from repeating the above process over all agents, and the last line follows from (F.6). For every $i \in [N]$, we have

$$\begin{aligned} &\left\langle \widetilde{r}_i^{(t)}, \pi_i^{(t+1)} - \pi_i^{(t)} \right\rangle + \tau \left(\mathcal{H}(\pi_i^{(t+1)}) - \mathcal{H}(\pi_i^{(t)})\right) \\ &= \left\langle r_i^{(t)}, \pi_i^{(t+1)} - \pi_i^{(t)} \right\rangle + \tau \left(\mathcal{H}(\pi_i^{(t+1)}) - \mathcal{H}(\pi_i^{(t)})\right) + \left\langle \widetilde{r}_i^{(t)} - r_i^{(t)}, \pi_i^{(t+1)} - \pi_i^{(t)} \right\rangle. \end{aligned}$$

We control the terms separately.

- For the first two terms, recall that taking logarithm on both sides of (7.5) gives

$$\eta r_i^{(t)} = \log \pi_i^{(t+1)} - (1 - \eta\tau) \log \pi_i^{(t)} + c\mathbf{1}$$

for some constant c. It follows that

$$\begin{aligned}
&\langle r_i^{(t)}, \pi_i^{(t+1)} - \pi_i^{(t)} \rangle + \tau \left(\mathcal{H}(\pi_i^{(t+1)}) - \mathcal{H}(\pi_i^{(t)}) \right) \\
&= \frac{1}{\eta} \langle \log \pi_i^{(t+1)} - \log \pi_i^{(t)}, \pi_i^{(t+1)} - \pi_i^{(t)} \rangle + \tau \left(\left\langle \log \pi_i^{(t)}, \pi_i^{(t+1)} - \pi_i^{(t)} \right\rangle + \mathcal{H}(\pi_i^{(t+1)}) - \mathcal{H}(\pi_i^{(t)}) \right) \\
&= \left(\frac{1}{\eta} - \tau \right) \mathsf{KL} \left(\pi_i^{(t+1)} \,\|\, \pi_i^{(t)} \right) + \frac{1}{\eta} \mathsf{KL} \left(\pi_i^{(t)} \,\|\, \pi_i^{(t+1)} \right). \qquad \text{(F.8)}
\end{aligned}$$

- For the third term, according to (F.4), we have

$$\left| \widetilde{r}_i^{(t)}(a) - r_i^{(t)}(a) \right| \le 2 d_{TV}(\widetilde{\pi}_{-i}^{(t)}, \pi_{-i}^{(t)}).$$

Hence,

$$\begin{aligned}
&\sum_{i=1}^N \left| \left\langle \widetilde{r}_i^{(t)} - r_i^{(t)}, \pi_i^{(t+1)} - \pi_i^{(t)} \right\rangle \right| \\
&\le 2 \sum_{i=1}^N d_{TV}(\widetilde{\pi}_{-i}^{(t)}, \pi_{-i}^{(t)}) \left\| \pi_i^{(t+1)} - \pi_i^{(t)} \right\|_1 \\
&\overset{\text{(i)}}{\le} \frac{2}{\sqrt{N}} \sum_{i=1}^N d_{TV}(\pi_{-i}^{(t)}, \widetilde{\pi}_{-i}^{(t)})^2 + \frac{\sqrt{N}}{2} \sum_{i=1}^N \left\| \pi_i^{(t+1)} - \pi_i^{(t)} \right\|_1^2 \\
&\overset{\text{(ii)}}{\le} \frac{1}{\sqrt{N}} \sum_{i=1}^N \mathsf{KL} \left(\pi_{-i}^{(t)} \,\|\, \widetilde{\pi}_{-i}^{(t)} \right) + \sqrt{N} \sum_{i=1}^N \mathsf{KL} \left(\pi_i^{(t+1)} \,\|\, \pi_i^{(t)} \right) \\
&\le \frac{1}{\sqrt{N}} \sum_{i=1}^N \mathsf{KL} \left(\pi^{(t)} \,\|\, \pi^{(t+1)} \right) + \sqrt{N} \mathsf{KL} \left(\pi^{(t+1)} \,\|\, \pi^{(t)} \right) \\
&= \sqrt{N} \left(\mathsf{KL} \left(\pi^{(t+1)} \,\|\, \pi^{(t)} \right) + \mathsf{KL} \left(\pi^{(t)} \,\|\, \pi^{(t+1)} \right) \right) = \sqrt{N} J(\pi^{(t+1)}, \pi^{(t)}),
\end{aligned}$$

where (i) results from Young's inequality and (ii) is due to Pinsker's inequality.

Combining all pieces together, we have

$$\begin{aligned}
\Phi_\tau^{(t+1)} - \Phi_\tau^{(t)} &\ge \left(\frac{1}{\eta} - \tau \right) \sum_{i=1}^N \left[\mathsf{KL} \left(\pi_i^{(t+1)} \,\|\, \pi_i^{(t)} \right) + \mathsf{KL} \left(\pi_i^{(t)} \,\|\, \pi_i^{(t+1)} \right) \right] - \sqrt{N} J(\pi^{(t+1)}, \pi^{(t)}) \\
&\ge \left(\frac{1}{\eta} - \sqrt{N} - \tau \right) J(\pi^{(t+1)}, \pi^{(t)}).
\end{aligned}$$

F.3.2 Proof of (F.5b)

Alternatively, we can decompose $\Phi_\tau^{(t+1)} - \Phi_\tau^{(t)}$ as

$$
\begin{aligned}
&\Phi_\tau^{(t+1)} - \Phi_\tau^{(t)} \\
&= \Phi^{(t+1)} - \Phi^{(t)} + \tau \sum_{i=1}^N \left[\mathcal{H}(\pi_i^{(t+1)}) - \mathcal{H}(\pi_i^{(t)})\right] \\
&= \sum_{i=1}^N \left[\Phi(\pi_i^{(t+1)}, \pi_{-i}^{(t)}) - \Phi^{(t)} + \tau\mathcal{H}(\pi_i^{(t+1)}) - \tau\mathcal{H}(\pi_i^{(t)})\right] + \Phi^{(t+1)} - \Phi^{(t)} - \sum_{i=1}^N \left[\Phi(\pi_i^{(t+1)}, \pi_{-i}^{(t)}) - \Phi^{(t)}\right] \\
&= \sum_{i=1}^N \left[u_i(\pi_i^{(t+1)}, \pi_{-i}^{(t)}) - u_i(\pi^{(t)}) + \tau\mathcal{H}(\pi_i^{(t+1)}) - \tau\mathcal{H}(\pi_i^{(t)})\right] + \Phi^{(t+1)} - \Phi^{(t)} - \sum_{i=1}^N \left[\Phi(\pi_i^{(t+1)}, \pi_{-i}^{(t)}) - \Phi^{(t)}\right].
\end{aligned}
$$

The first term is lower bounded by $(1/\eta - \tau)J(\pi_i^{(t+1)}, \pi_i^{(t)})$ as shown in (F.8). For the remaining terms, we have

$$
\begin{aligned}
&\left|\Phi^{(t+1)} - \Phi^{(t)} - \sum_{i=1}^N \left[\Phi(\pi_i^{(t+1)}, \pi_{-i}^{(t)}) - \Phi^{(t)}\right]\right| \\
&\le \sum_{\boldsymbol{a}\in\mathcal{A}^N} \Phi(\boldsymbol{a})\pi^{(t)}(\boldsymbol{a}) \left|\frac{\pi^{(t+1)}(\boldsymbol{a})}{\pi^{(t)}(\boldsymbol{a})} - 1 - \sum_{i=1}^N \left(\frac{\pi_i^{(t+1)}(a_i)}{\pi_i^{(t)}(a_i)} - 1\right)\right|. \qquad \text{(F.9)}
\end{aligned}
$$

To continue, we need the following elementary lemma, which will be proved at the end.

Lemma 47. *For all $x \in (-1, \infty)$, it holds that*

$$0 \le x - \log(1+x) \le x\log(1+x).$$

Invoking Lemma 47 to obtain

$$
\begin{aligned}
&\left|\frac{\pi^{(t+1)}(\boldsymbol{a})}{\pi^{(t)}(\boldsymbol{a})} - 1 - \sum_{i=1}^N \left(\frac{\pi_i^{(t+1)}(a_i)}{\pi_i^{(t)}(a_i)} - 1\right)\right| \\
&= \left|\frac{\pi^{(t+1)}(\boldsymbol{a})}{\pi^{(t)}(\boldsymbol{a})} - 1 - \log\frac{\pi^{(t+1)}(\boldsymbol{a})}{\pi^{(t)}(\boldsymbol{a})} - \sum_{i=1}^N \left(\frac{\pi_i^{(t+1)}(a_i)}{\pi_i^{(t)}(a_i)} - 1 - \log\frac{\pi_i^{(t+1)}(a_i)}{\pi_i^{(t)}(a_i)}\right)\right| \\
&\le \left(\frac{\pi^{(t+1)}(\boldsymbol{a})}{\pi^{(t)}(\boldsymbol{a})} - 1\right)\log\frac{\pi^{(t+1)}(\boldsymbol{a})}{\pi^{(t)}(\boldsymbol{a})} + \sum_{i=1}^N \left(\frac{\pi_i^{(t+1)}(a_i)}{\pi_i^{(t)}(a_i)} - 1\right)\log\frac{\pi_i^{(t+1)}(a_i)}{\pi_i^{(t)}(a_i)}.
\end{aligned}
$$

Plugging the above inequality into (F.9) yields

$$
\begin{aligned}
&\left|\Phi^{(t+1)} - \Phi^{(t)} - \sum_{i=1}^N \left[\Phi(\pi_i^{(t+1)}, \pi_{-i}^{(t)}) - \Phi^{(t)}\right]\right| \\
&\le \Phi_{\max} \sum_{\boldsymbol{a}\in\mathcal{A}^N} \pi^{(t)}(\boldsymbol{a}) \left[\left(\frac{\pi^{(t+1)}(\boldsymbol{a})}{\pi^{(t)}(\boldsymbol{a})} - 1\right)\log\frac{\pi^{(t+1)}(\boldsymbol{a})}{\pi^{(t)}(\boldsymbol{a})} + \sum_{i=1}^N \left(\frac{\pi_i^{(t+1)}(a_i)}{\pi_i^{(t)}(a_i)} - 1\right)\log\frac{\pi_i^{(t+1)}(a_i)}{\pi_i^{(t)}(a_i)}\right] \\
&= \Phi_{\max} \sum_{\boldsymbol{a}\in\mathcal{A}^N} \left[\left(\pi^{(t+1)}(\boldsymbol{a}) - \pi^{(t)}(\boldsymbol{a})\right)\log\frac{\pi^{(t+1)}(\boldsymbol{a})}{\pi^{(t)}(\boldsymbol{a})}\right] + \Phi_{\max} \sum_{i=1}^N \sum_{a_i\in\mathcal{A}} \left(\pi_i^{(t+1)}(a_i) - \pi_i^{(t)}(a_i)\right)\log\frac{\pi_i^{(t+1)}(a_i)}{\pi_i^{(t)}(a_i)} \\
&= \Phi_{\max}\left(J(\pi^{(t+1)}, \pi^{(t)}) + \sum_{i=1}^N J(\pi_i^{(t+1)}, \pi_i^{(t)})\right) = 2\Phi_{\max} J(\pi^{(t+1)}, \pi^{(t)}).
\end{aligned}
$$

Combining all pieces together, we have

$$\Phi_\tau^{(t+1)} - \Phi_\tau^{(t)} \geq \left(\frac{1}{\eta} - 2\Phi_{\max} - \tau\right) \sum_{i=1}^{N} J(\pi^{(t+1)}, \pi^{(t)}).$$

Proof of Lemma 47. We have $x - \log(1+x) = x\log(1+x) = 0$ and $(x - \log(1+x))' = (x\log(1+x))' = 0$ when $x = 0$. It follows that $x - \log(1+x) \geq 0$ since log is concave. With straightforward calculation, we get

$$(x\log(1+x))' - (x - \log(1+x))' = \log(1+x) \begin{cases} \geq 0 & x \geq 0 \\ < 0 & -1 < x < 0 \end{cases},$$

which implies $x - \log(1+x) \leq x\log(1+x)$. □

F.3.3 Proof of Corollary 1

By noting that $\pi_i^{\star(0)} \propto \exp\left(r_i^{(0)}/\tau\right)$, and with uniform policy initialization $\pi_i^{(0)} \propto 1$, we can conclude

$$\left\|\log \pi_i^{(0)} - \log \pi_i^{\star(0)}\right\|_\infty \leq 2 \left\|\frac{r_i^{(t)}}{\tau} - 0\right\|_\infty \leq \frac{2}{\tau}.$$

where the first inequality follows from Lemma 15, and the second inequality is true since the payoff is bounded by 1. On the other hand, we have

$$\begin{aligned} \Phi_\tau^{(T)} - \Phi_\tau^{(0)} &= \Phi(\pi^{(T)}) - \Phi(\pi^{(0)}) + \tau\mathcal{H}(\pi^{(T)}) - \tau\mathcal{H}(\pi^{(0)}) \\ &\leq \Phi(\pi^{(T)}) - \Phi(\pi^{(0)}) \leq \Phi_{\max}, \end{aligned}$$

where the first inequality uses the fact that the entropy is maximized for uniform policies, and the second inequality uses $0 \leq \Phi(\pi) \leq \Phi_{\max}$ for any π. Combining the above two bounds with Theorem 12, we have

$$\frac{1}{T}\sum_{t=1}^{T} \texttt{QRE-gap}_\tau^{(t)} \leq \frac{4}{\tau\eta T} + \frac{2}{\tau}\sqrt{\frac{2\Phi_{\max}}{\eta T}}.$$

Setting $\eta = \frac{1}{2(\min\{\sqrt{N}, 2\Phi_{\max}\}+\tau)}$ and $T = \mathcal{O}\left(\frac{\min\{\sqrt{N},\Phi_{\max}\}\Phi_{\max}}{\tau^2\varepsilon^2}\right)$ thus completes the proof.

Appendix G

Proofs for Chapter 8

G.1 Analysis for the online setting

G.1.1 Proof of Theorem 13

For ease of presentation, we assume that $\mathcal{R}$ is finite, i.e., $|\mathcal{R}| < \infty$. The general case can be directly obtained using a covering number argument, which we refer to [Liu et al., 2024a, Jin et al., 2022] for interested readers.

We start by decomposing the regret into two parts:

$$\begin{aligned}\mathsf{Regret} &:= \sum_{t=1}^{T} \left[J^\star(r^\star) - J(r^\star, \pi^{(t)})\right] \\ &= \underbrace{\sum_{t=1}^{T} \left[J^\star(r^\star) - J^\star(r^{(t)})\right]}_{\textbf{Term (i)}} + \underbrace{\sum_{t=1}^{T} \left[J(r^{(t)}, \pi^{(t)}) - J(r^\star, \pi^{(t)})\right]}_{\textbf{Term (ii)}}. \end{aligned} \tag{G.1}$$

Step 1: bounding term (i). By the choice of $r^{(t)}$, we have

$$\ell(r^{(t)}, \mathcal{D}^{(t-1)}) - \alpha J^\star(r^{(t)}) \le \ell(r^\star, \mathcal{D}^{(t-1)}) - \alpha J^\star(r^\star). \tag{G.2}$$

Rearranging terms,

$$J^\star(r^\star) - J^\star(r^{(t)}) \le \frac{1}{\alpha}\left[\ell(r^\star, \mathcal{D}^{(t-1)}) - \ell(r^{(t)}, \mathcal{D}^{(t-1)})\right]. \tag{G.3}$$

The following lemma is adapted from [Liu et al., 2024a, Proposition 5.3], whose proof is deferred to Appendix G.1.2.

Lemma 48. *Let $\delta \in (0,1)$. With probability $1-\delta$, we have*

$$\begin{aligned}&\ell(r^\star, \mathcal{D}^{(t-1)}) - \ell(r^{(t)}, \mathcal{D}^{(t-1)}) \\ &\le -2\sum_{s=1}^{t-1} \mathop{\mathbb{E}}_{\substack{x\sim\rho,\\ (y_1,y_2)\sim\pi^{(s)}(\cdot|x)}} \left[D_H^2(\mathbb{P}_{r^{(t)}}(\cdot|x,y_1,y_2) \,\|\, \mathbb{P}_{r^\star}(\cdot|x,y_1,y_2))\right] + 2\log(|\mathcal{R}|/\delta). \end{aligned} \tag{G.4}$$

Here, $D_H(\cdot\|\cdot)$ is the Hellinger distance, $\mathbb{P}_r(\cdot|x,y_1,y_2)$ denotes the Bernoulli distribution of the comparison result of (x,y_1) and (x,y_2) under reward model r.

Putting the above inequalities together, it holds with probability $1-\delta$ that

$$\textbf{Term (i)} \le -\frac{2}{\alpha}\sum_{t=1}^{T}\sum_{s=1}^{t-1} \mathop{\mathbb{E}}_{\substack{x^{(s)}\sim\rho,\\ (y_1^{(s)},y_2^{(s)})\sim\pi^{(s)}(\cdot|x^{(s)})}} \left[D_{\mathrm{H}}^2(\mathbb{P}_{r^{(t)}}(\cdot|x^{(s)},y_1^{(s)},y_2^{(s)}) \,\|\, \mathbb{P}_{r^\star}(\cdot|x^{(s)},y_1^{(s)},y_2^{(s)}))\right] + 2\alpha^{-1}T\log(|\mathcal{R}|/\delta). \tag{G.5}$$

Step 2: breaking down term (ii) with the elliptical potential lemma. The linear function approximation form (8.18) allows us to write

$$\mathop{\mathbb{E}}_{x\sim\rho,y\sim\pi_{r_2}(\cdot|x)}[r_1(x,y)-r^\star(x,y)] = \langle W(r_1), X(r_2)\rangle\,, \tag{G.6}$$

where $X, W : \mathcal{R}\to\mathbb{R}^d$ is given by

$$X(r_\theta) = 2C \mathop{\mathbb{E}}_{x\sim\rho,y\sim\pi_{r_\theta}(\cdot|x)}[\phi(x,y)]\,, \qquad W(r_\theta) = \frac{\theta-\theta^\star}{2C}. \tag{G.7}$$

Let

$$\Sigma_t = \varepsilon I + \sum_{s=1}^{t-1} X(r^{(t)})X(r^{(t)})^\top \tag{G.8}$$

for some $\varepsilon>0$. We begin by decomposing term (ii) as

$$\begin{aligned}
\textbf{Term (ii)} &= \sum_{t=1}^{T} \mathop{\mathbb{E}}_{x\sim\rho,y\sim\pi^{(t)}(\cdot|x)}\left[r^{(t)}(x,y)-r^\star(x,y)\right] \\
&= \sum_{t=1}^{T}\left\langle W(r^{(t)}), X(r^{(t)})\right\rangle \\
&= \sum_{t=1}^{T}\left\langle W(r^{(t)}), X(r^{(t)})\right\rangle \mathbf{1}\{\|X(r^{(t)})\|_{\Sigma_t^{-1}}\le 1\} \\
&\qquad + \sum_{t=1}^{T}\left\langle W(r^{(t)}), X(r^{(t)})\right\rangle \mathbf{1}\{\|X(r^{(t)})\|_{\Sigma_t^{-1}}> 1\},
\end{aligned} \tag{G.9}$$

where $\mathbf{1}\{A\}$ is an indicator function of event A. To proceed, we recall the elliptical potential lemma for controlling the cumulative sum of $\min\{\|X(r^{(t)})\|^2_{\Sigma_t^{-1}},1\}$.

Lemma 49 ([Abbasi-Yadkori et al., 2011, Lemma 11])**.** *Let $\{X_t\}$ be a sequence in $\mathbb{R}^d$ and $\Lambda_0\in\mathbb{R}^{d\times d}$ a positive definite matrix. Define $\Lambda_t=\Lambda_0+\sum_{s=1}^t X_sX_s^\top$. Assume $\|X_t\|\le L$ for all t. It holds that*

$$\begin{aligned}
\sum_{t=1}^{T}\min\{\|X_t\|^2_{\Lambda_t^{-1}},1\} &\le 2\log\left(\frac{det(\Lambda_T)}{det(\Lambda_0)}\right) \\
&\le 2(d\log((trace(\Lambda_0)+TL^2)/d) - \log det(\Lambda_0)).
\end{aligned}$$

Applying the above lemma yields

$$\sum_{t=1}^{T}\min\{\|X(r^{(t)})\|^2_{\Sigma_t^{-1}},1\} \le \min\left\{2d\log\left(\frac{4C^4T/d+\varepsilon}{\varepsilon}\right), T\right\} := d(\varepsilon). \tag{G.10}$$

We now control the two terms in (G.9). • The first term of (G.9) can be bounded by

$$
\begin{aligned}
&\sum_{t=1}^{T}\Big\langle W(r^{(t)}), X(r^{(t)})\Big\rangle \mathbf{1}\{\|X(r^{(t)})\|_{\Sigma_t^{-1}} \le 1\} \\
&\le \sum_{t=1}^{T}\|W(r^{(t)})\|_{\Sigma_t}\|X(r^{(t)})\|_{\Sigma_t^{-1}}\mathbf{1}\{\|X(r^{(t)})\|_{\Sigma_t^{-1}} \le 1\} \\
&\le \sum_{t=1}^{T}\|W(r^{(t)})\|_{\Sigma_t}\min\Big\{\|X(r^{(t)})\|_{\Sigma_t^{-1}}, 1\Big\} \\
&= \sum_{t=1}^{T}\Big[\varepsilon\|W(r^{(t)})\|_2^2 + \sum_{s=1}^{t-1}\Big\langle W(r^{(t)}), X(r^{(s)})\Big\rangle^2\Big]^{1/2}\min\Big\{\|X(r^{(t)})\|_{\Sigma_t^{-1}}^2, 1\Big\}^{1/2} \\
&\overset{\text{(i)}}{\le} \Big\{\sum_{t=1}^{T}\Big[\varepsilon\|W(r^{(t)})\|_2^2 + \sum_{s=1}^{t-1}\Big\langle W(r^{(t)}), X(r^{(s)})\Big\rangle^2\Big]\Big\}^{1/2}\Big\{\sum_{t=1}^{T}\min\Big\{\|X(r^{(t)})\|_{\Sigma_t^{-1}}^2, 1\Big\}\Big\}^{1/2} \\
&\overset{\text{(ii)}}{\le} \sqrt{d(\varepsilon)}\Big\{\sum_{t=1}^{T}\sum_{s=1}^{t-1}\Big\langle W(r^{(t)}), X(r^{(s)})\Big\rangle^2\Big\}^{1/2} + \sqrt{d(\varepsilon)\varepsilon T} \\
&\overset{\text{(iii)}}{\le} \frac{d(\varepsilon)}{2\mu} + \frac{\mu}{2}\sum_{t=1}^{T}\sum_{s=1}^{t-1}\Big\langle W(r^{(t)}), X(r^{(s)})\Big\rangle^2 + \sqrt{d(\varepsilon)\varepsilon T}.
\end{aligned}
\tag{G.11}
$$

Here, (i) is due to Cauchy-Schwarz inequality, (ii) is due to $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$ for $\forall a, b \ge 0$, and (iii) results from Young's inequality. We leave the constant $\mu > 0$ to be determined later. • The second term of (G.9) can be bounded by

$$
\begin{aligned}
\sum_{t=1}^{T}\Big\langle W(r^{(t)}), X(r^{(t)})\Big\rangle \mathbf{1}\{\|X(r^{(t)})\|_{\Sigma_t^{-1}} > 1\} &\le C\sum_{t=1}^{T}\mathbf{1}\{\|X(r^{(t)})\|_{\Sigma_t^{-1}} > 1\} \\
&\le C\sum_{t=1}^{T}\min\{\|X(r^{(t)})\|_{\Sigma_t^{-1}}^2, 1\} \le Cd(\varepsilon),
\end{aligned}
\tag{G.12}
$$

where the first inequality follows from $\|X(r^{(t)})\|_2 \le 2C$ and $\|W(r^{(t)})\|_2 \le 1/2$ since $\|\phi(x,y)\|_2 \le 1$. Putting (G.9), (G.11) and (G.12) together, we arrive at

$$
\textbf{Term (ii)} \le \frac{d(\varepsilon)}{2\mu} + \frac{\mu}{2}\sum_{t=1}^{T}\sum_{s=1}^{t-1}\Big\langle W(r^{(t)}), X(r^{(s)})\Big\rangle^2 + \sqrt{d(\varepsilon)\varepsilon T} + Cd(\varepsilon). \tag{G.13}
$$

Step 3: continuing bounding term (ii). It boils down to control $\big\langle W(r^{(t)}), X(r^{(s)})\big\rangle^2$. We have

$$
\begin{aligned}
\Big\langle W(r^{(t)}), X(r^{(s)})\Big\rangle &= \mathop{\mathbb{E}}_{\substack{x\sim\rho,\\ y\sim\pi^{(s)}(\cdot|x)}}\Big[r^{(t)}(x,y) - r^\star(x,y)\Big] \\
&= \mathop{\mathbb{E}}_{\substack{x\sim\rho,\\ y_1\sim\pi^{(s)}(\cdot|x)}}\Big[r^{(t)}(x,y_1) - r^\star(x,y_1)\Big] - \mathop{\mathbb{E}}_{\substack{x\sim\rho,\\ y_2\sim\pi_{\mathrm{cal}}(\cdot|x)}}\Big[r^{(t)}(x,y_2) - r^\star(x,y_2)\Big] \\
&= \mathop{\mathbb{E}}_{\substack{x\sim\rho,\\ y_1\sim\pi^{(s)}(\cdot|x),\\ y_2\sim\pi_{\mathrm{cal}}(\cdot|x)}}\Big[\delta_x(r^{(t)}, r^\star, y_1, y_2)\Big],
\end{aligned}
\tag{G.14}
$$

where $\delta_x(r_1, r_2, y_1, y_2) \coloneqq r_1(x, y_1) - r_1(x, y_2) - (r_2(x, y_1) - r_2(x, y_2))$. Therefore,

$$\begin{aligned}
\left\langle W(r^{(t)}), X(r^{(s)}) \right\rangle^2 &= \mathop{\mathbb{E}}_{\substack{x\sim\rho,\\ y_1\sim\pi^{(s)}(\cdot|x),\\ y_2\sim\pi_{\mathrm{cal}}(\cdot|x)}} \left[\delta_x(r^{(t)}, r^\star, y_1, y_2)\right]^2 \\
&= \mathop{\mathbb{E}}_{\substack{x\sim\rho,\\ y_1\sim\pi^{(s)}(\cdot|x),\\ y_2\sim\pi_{\mathrm{cal}}(\cdot|x)}} \left[\delta_x(r^{(t)}, r^\star, y_1, y_2)^2\right] - \mathop{\mathrm{Var}}_{\substack{x\sim\rho,\\ y_1\sim\pi^{(s)}(\cdot|x),\\ y_2\sim\pi_{\mathrm{cal}}(\cdot|x)}} [\delta_x(r^{(t)}, r^\star, y_1, y_2)^2] \\
&\le \mathop{\mathbb{E}}_{\substack{x\sim\rho,\\ y_1\sim\pi^{(s)}(\cdot|x),\\ y_2\sim\pi_{\mathrm{cal}}(\cdot|x)}} \left[\delta_x(r^{(t)}, r^\star, y_1, y_2)^2\right] \\
&\le \sup_{x,y} \frac{\pi_{\mathrm{cal}}(y|x)}{\pi^{(s)}(y|x)} \cdot \mathop{\mathbb{E}}_{\substack{x\sim\rho,\\ y_1,y_2\sim\pi^{(s)}(\cdot|x)}} \left[\delta_x(r^{(t)}, r^\star, y_1, y_2)^2\right] \\
&\le \sup_{x,y} \frac{\pi_{\mathrm{ref}}(y|x)}{\pi^{(s)}(y|x)} \cdot \sup_{x,y} \frac{\pi_{\mathrm{cal}}(y|x)}{\pi_{\mathrm{ref}}(y|x)} \cdot \mathop{\mathbb{E}}_{\substack{x\sim\rho,\\ y_1,y_2\sim\pi^{(s)}(\cdot|x)}} \left[\delta_x(r^{(t)}, r^\star, y_1, y_2)^2\right]. \qquad \text{(G.15)}
\end{aligned}$$

Recall from (8.6) that $\pi^{(s)}(y|x) \propto \pi_{\mathrm{ref}}(y|x)\exp(r^{(s)}(x,y)/\beta)$. It follows that $|\log \pi^{(s)}(y|x) - \log \pi_{\mathrm{ref}}(y|x)| \le 2\|r^{(s)}(x,\cdot)\|_\infty \le 2C/\beta$ (see *e.g.*, [Cen et al., 2022b, Appendix A.2]), and hence $\sup_{x,y} \frac{\pi_{\mathrm{ref}}(y|x)}{\pi^{(s)}(y|x)} \le \exp(2C/\beta)$. To proceed, we demonstrate in the following lemma that δ^2 can be upper bounded by the corresponding Hellinger distance, whose proof is deferred to Appendix G.1.3.

Lemma 50. *Assume bounded reward $\|r_1\|_\infty \le C$, $\|r_2\|_\infty \le C$. We have*

$$\delta_x(r_1, r_2, y_1, y_2)^2 \le 2(3 + \exp(2C))^2 D_H^2(\mathbb{P}_{r_1}(\cdot|x, y_1, y_2) \,\|\, \mathbb{P}_{r_2}(\cdot|x, y_1, y_2)).$$

With the above lemma we arrive at

$$\begin{aligned}
&\left\langle W(r^{(t)}), X(r^{(s)}) \right\rangle^2 \\
&\le 2(3 + \exp(2C))^2 \exp(2C/\beta)\kappa \cdot \mathop{\mathbb{E}}_{\substack{x\sim\rho,\\ y_1,y_2\sim\pi^{(s)}(\cdot|x)}} \left[D_{\mathrm{H}}^2(\mathbb{P}_{r^{(t)}}(\cdot|x, y_1, y_2) \,\|\, \mathbb{P}_{r^\star}(\cdot|x, y_1, y_2))\right].
\end{aligned}$$

where we denote $\kappa = \sup_{x,y} \frac{\pi_{\mathrm{cal}}(y|x)}{\pi_{\mathrm{ref}}(y|x)}$. Plugging the above bound into (G.13), we get

Term (ii)

$$\begin{aligned}
&\le \frac{d(\varepsilon)}{2\mu} + \mu(3 + \exp(2C))^2 \exp(2C/\beta)\kappa \cdot \sum_{t=1}^{T}\sum_{s=1}^{t-1} \mathop{\mathbb{E}}_{\substack{x\sim\rho,\\ y_1,y_2\sim\pi^{(s)}(\cdot|x)}} \left[D_{\mathrm{H}}^2(\mathbb{P}_{r^{(t)}}(\cdot|x, y_1, y_2) \,\|\, \mathbb{P}_{r^\star}(\cdot|x, y_1, y_2))\right] \\
&\quad + 2B\sqrt{d(\varepsilon)\varepsilon T} + Cd(\varepsilon). \qquad \text{(G.16)}
\end{aligned}$$

Step 4: finishing up. Combining (G.1), (G.5) and (G.16), with probability $1-\delta$ we have

$$\mathsf{Regret} \le \frac{2T\log(|\mathcal{R}|/\delta)}{\alpha} + \frac{d(\varepsilon)}{2\mu} + \sqrt{d(\varepsilon)\varepsilon T} + Cd(\varepsilon) \qquad \text{(G.17)}$$

as long as $\alpha\mu(3+\exp(2C))^2\exp(2C/\beta)\kappa \le 2$. Setting $\alpha \asymp \frac{1}{\exp(2C+C/\beta)}\sqrt{\frac{T}{\kappa d(\varepsilon)}}$, $\mu \asymp \frac{1}{\exp(2C+C/\beta)}\sqrt{\frac{d(\varepsilon)}{\kappa T}}$, and $\varepsilon = 1$, we arrive at

$$\mathsf{Regret} \le \widetilde{\mathcal{O}}((\exp(2C + C/\beta))\sqrt{\kappa d T})$$

as claimed.

G.1.2 Proof of Lemma 48

To begin, we have

$$\ell(r^\star, \mathcal{D}^{(t-1)}) - \ell(r^{(t)}, \mathcal{D}^{(t-1)}) = -\log\frac{\mathbb{P}(\mathcal{D}^{(t-1)}|r^\star)}{\mathbb{P}(\mathcal{D}^{(t-1)}|r^{(t)})} = -\sum_{s=1}^{t-1} X^s_{r^{(t)}}, \tag{G.18}$$

where we denote

$$X^s_r = \log\frac{\mathbb{P}_{r^\star}(y^{(s)}_+ \succ y^{(s)}_-|x^{(s)})}{\mathbb{P}_r(y^{(s)}_+ \succ y^{(s)}_-|x^{(s)})}. \tag{G.19}$$

To proceed, we recall a useful martingale exponential inequality.

Lemma 51 ([Zhang, 2023, Theorem 13.2],[Liu et al., 2024a, Lemma D.1])**.** *Let $\{X_t\}_{t=1}^\infty$ be a sequence of real-valued random variables adapted to filtration $\{\mathcal{F}_t\}_{t=1}^\infty$. It holds with probability $1-\delta$ such that for any $t \ge 1$,*

$$-\sum_{s=1}^{t} X_s \le \sum_{s=1}^{t} \log\mathbb{E}\left[\exp(-X_s)|\mathcal{F}_{s-1}\right] + \log(1/\delta).$$

Applying the above lemma to $\{\frac{1}{2}X^t_r\}_{t=1}^\infty$ along with the filtration $\{\mathcal{F}_t\}_{t=1}^\infty$ with $\mathcal{F}_t$ given by the σ-algebra of $\{(x^{(s)}, y^{(s)}_+, y^{(s)}_-) : s \le t\}$, we conclude that it holds with probability $1-\delta$ that

$$\begin{aligned}
-\frac{1}{2}\sum_{s=1}^{t-1} X^s_r &\le \sum_{s=1}^{t-1} \log\mathbb{E}\left[\exp\Big\{-\frac{1}{2}X^s_r\Big\}\Big|\mathcal{F}_{s-1}\right] + \log(|\mathcal{R}|/\delta) \\
&\le \sum_{s=1}^{t-1}\left(\mathbb{E}\left[\exp\Big\{-\frac{1}{2}X^s_r\Big\}\Big|\mathcal{F}_{s-1}\right] - 1\right) + \log(|\mathcal{R}|/\delta),
\end{aligned} \tag{G.20}$$

where the last step results from the inequality $\log(1+x) \le x$ for all $x \ge -1$. To proceed, note that

$$
\begin{aligned}
&\mathbb{E}\left[\exp\Big\{-\frac{1}{2}X_r^s\Big\}\Big|\mathcal{F}_{s-1}\right] \\
&= \mathbb{E}\left[\sqrt{\frac{\mathbb{P}_r(y_+^{(s)} \succ y_-^{(s)}|x^{(s)})}{\mathbb{P}_{r^\star}(y_+^{(s)} \succ y_-^{(s)}|x^{(s)})}}\,\Big|\,\mathcal{F}_{s-1}\right] \\
&= \mathop{\mathbb{E}}_{\substack{x^{(s)}\sim\rho,\\ (y_1^{(s)},y_2^{(s)})\sim\pi^{(s)}(\cdot|x^{(s)}),\\ (+,-)\sim\mathbb{P}_{r^\star}}}\left[\sqrt{\frac{\mathbb{P}_r(y_+^{(s)} \succ y_-^{(s)}|x^{(s)})}{\mathbb{P}_{r^\star}(y_+^{(s)} \succ y_-^{(s)}|x^{(s)})}}\right] \\
&= \mathop{\mathbb{E}}_{\substack{x^{(s)}\sim\rho,\\ (y_1^{(s)},y_2^{(s)})\sim\pi^{(s)}(\cdot|x^{(s)})}}\left[\sum_{(+,-)}\sqrt{\mathbb{P}_r(y_+^{(s)} \succ y_-^{(s)}|x^{(s)})\cdot\mathbb{P}_{r^\star}(y_+^{(s)} \succ y_-^{(s)}|x^{(s)})}\right] \\
&= 1 - \frac{1}{2}\mathop{\mathbb{E}}_{\substack{x^{(s)}\sim\rho,\\ (y_1^{(s)},y_2^{(s)})\sim\pi^{(s)}(\cdot|x^{(s)})}}\left[\sum_{(+,-)}\Big(\sqrt{\mathbb{P}_r(y_+^{(s)} \succ y_-^{(s)}|x^{(s)})} - \sqrt{\mathbb{P}_{r^\star}(y_+^{(s)} \succ y_-^{(s)}|x^{(s)})}\Big)^2\right] \\
&= 1 - \mathop{\mathbb{E}}_{\substack{x\sim\rho,\\ (y_1,y_2)\sim\pi^{(s)}(\cdot|x)}}\left[D_{\mathrm{H}}^2(\mathbb{P}_r(\cdot|x,y_1,y_2)\,\|\,\mathbb{P}_{r^\star}(\cdot|x,y_1,y_2)\right],
\end{aligned}
$$

where we denote by $\sum_{(+,-)}$ the summation over different comparison results. Plugging the above equality into (G.20) completes the proof.

G.1.3 Proof of Lemma 50

By the mean value theorem, we have

$$
\begin{aligned}
\big|\mathbb{P}_{r_1}(y_1 \succ y_2|x) - \mathbb{P}_{r_2}(y_1 \succ y_2|x)\big| &= \big|\sigma(r_1(x,y_1)-r_1(x,y_2)) - \sigma(r_2(x,y_1)-r_2(x,y_2))\big| \\
&= \big|\delta_x(r_1,r_2,y_1,y_2)\cdot\sigma'(\xi)\big| \\
&= \big|\delta_x(r_1,r_2,y_1,y_2)\big|\cdot\sigma(\xi)(1-\sigma(\xi))
\end{aligned}
$$

for some ξ between $r_1(x,y_1)-r_1(x,y_2)$ and $r_2(x,y_1)-r_2(x,y_2)$. Since $|\xi| \le 2C$, we have

$$
\sigma(\xi)(1-\sigma(\xi)) \ge \sigma(2C)(1-\sigma(2C)) \ge \frac{1}{3+\exp(2C)}. \tag{G.21}
$$

Putting together,

$$
\begin{aligned}
\big|\delta_x(r_1,r_2,y_1,y_2)\big| &\le (3+\exp(2C))\big|\mathbb{P}_{r_1}(y_1 \succ y_2|x) - \mathbb{P}_{r_2}(y_1 \succ y_2|x)\big| \\
&= (3+\exp(2C))\mathrm{TV}(\mathbb{P}_{r_1}(\cdot|x,y_1,y_2),\mathbb{P}_{r_2}(\cdot|x,y_1,y_2)) \\
&\le (3+\exp(2C))\sqrt{2}D_{\mathrm{H}}(\mathbb{P}_{r_1}(\cdot|x,y_1,y_2)\,\|\,\mathbb{P}_{r_2}(\cdot|x,y_1,y_2)).
\end{aligned}
$$

G.2 Analysis for the offline setting

G.2.1 Proof of Lemma 2

By definition, the objective function $\ell(r,\mathcal{D})+\alpha J(r,\pi)$ is strongly concave over π, and convex over r. By Danskin's theorem, we have

$$\nabla_r\big(\max_{\pi}[\ell(\widehat{r},\mathcal{D})+\alpha J(\widehat{r},\pi)]\big)=\nabla_r\big(\ell(\widehat{r},\mathcal{D})+\alpha J(\widehat{r},\widehat{\pi})\big).$$

Therefore, for any r', by convexity of the objective function we have

$$\begin{aligned}\ell(r',\mathcal{D})+\alpha J(r',\widehat{\pi}) &\geq \ell(\widehat{r},\mathcal{D})+\alpha J(\widehat{r},\widehat{\pi})+\big\langle r'-\widehat{r},\nabla_r\big(\ell(\widehat{r},\mathcal{D})+\alpha J(\widehat{r},\widehat{\pi})\big)\big\rangle\\ &=\ell(\widehat{r},\mathcal{D})+\alpha J(\widehat{r},\widehat{\pi})+\Big\langle r'-\widehat{r},\nabla_r\big(\max_{\pi}[\ell(\widehat{r},\mathcal{D})+\alpha J(\widehat{r},\pi)]\big)\Big\rangle\\ &\geq \ell(\widehat{r},\mathcal{D})+\alpha J(\widehat{r},\widehat{\pi}).\end{aligned}$$

The last line is due to the definition of $\widehat{r}$ (c.f. (8.23)). The other relation, $\ell(\widehat{r},\mathcal{D})+\alpha J(\widehat{r},\widehat{\pi})\geq \ell(\widehat{r},\mathcal{D})+\alpha J(\widehat{r},\pi')$, follows directly from the definition of $\widehat{\pi}$ (c.f. (8.20)).

G.2.2 Proof of Theorem 14

We decompose the sub-optimality gap of $\widehat{\pi}$ by

$$\begin{aligned}&J^\star(r^\star)-J(r^\star,\widehat{\pi})\\ &=\big[J(r^\star,\pi^\star)-J(\widehat{r},\pi^\star)\big]+\big[J(\widehat{r},\pi^\star)-J(\widehat{r},\widehat{\pi})\big]+\big[J(\widehat{r},\widehat{\pi})-J(r^\star,\widehat{\pi})\big]\\ &\leq \underbrace{\big[J(r^\star,\pi^\star)-J(\widehat{r},\pi^\star)\big]}_{\textbf{Term (i)}}+\underbrace{\big[J(\widehat{r},\widehat{\pi})-J(r^\star,\widehat{\pi})\big]}_{\textbf{Term (ii)}},\end{aligned}\tag{G.22}$$

where the last line is due to $J(\widehat{r},\pi^\star)\leq J(\widehat{r},\widehat{\pi})$ according to the definition of $\widehat{\pi}$ (c.f. (8.20)). We proceed to bound the two terms separately. Here we have written $\widehat{r}=r_{\widehat{\theta}}$ for notational simplicity. In addition, we denote the MLE estimate by $r_{\mathsf{MLE}}=r_{\theta_{\mathsf{MLE}}}$.

By the definition of $J(r,\pi)$ (cf. (8.4)), it follows that term (i) in (G.22) can be further decomposed as

$$\begin{aligned}\textbf{Term (i)} &= \mathop{\mathbb{E}}_{\substack{x\sim\rho,\\ y\sim\pi^\star(\cdot|x)}}[r^\star(x,y)-\widehat{r}(x,y)]\\ &= \mathop{\mathbb{E}}_{\substack{x\sim\rho,\\ y\sim\pi^\star(\cdot|x)}}\Big[\Big\langle\phi(x,y),\theta^\star-\widehat{\theta}\Big\rangle\Big]\\ &= \underbrace{\mathop{\mathbb{E}}_{\substack{x\sim\rho,\\ y\sim\pi^\star(\cdot|x)}}[\langle\phi(x,y),\theta^\star-\theta_{\mathsf{MLE}}\rangle]}_{\textbf{Term (ia)}}+\underbrace{\mathop{\mathbb{E}}_{\substack{x\sim\rho,\\ y\sim\pi^\star(\cdot|x)}}\Big[\Big\langle\phi(x,y),\theta_{\mathsf{MLE}}-\widehat{\theta}\Big\rangle\Big]}_{\textbf{Term (ib)}},\end{aligned}\tag{G.23}$$

where $r_{\mathsf{MLE}}(x,y)=\langle\phi(x,y),\theta_{\mathsf{MLE}}\rangle$.

Step 1: bounding term (ia). To continue, we recall a useful lemma from [Zhu et al., 2023].

Lemma 52 ([Zhu et al., 2023, Lemma 3.1])**.** *For any $\lambda>0$ and $\delta\in(0,1)$, with probability at least $1-\delta$,*

$$\|\theta_{\mathsf{MLE}}-\theta^\star\|_{\Sigma_{\mathcal{D}}+\lambda I}\leq\mathcal{O}\left((3+\exp(C))\sqrt{\frac{d+\log(1/\delta)}{N}}+\sqrt{\lambda C^2}\right).$$

In addition, we have

$$\frac{1}{3+\exp(C)}\Sigma_{\mathcal{D}} \preceq \frac{1}{N}\nabla_\theta^2 \ell(r_\theta, \mathcal{D}) \preceq \frac{1}{4}\Sigma_{\mathcal{D}} \tag{G.24}$$

for all θ such that $\|r_\theta\|_\infty \le C$.

The first term of (G.23) can be bounded with Lemma 52 as

$$\begin{aligned}\textbf{Term (ia)} &\le \|\theta^\star - \theta_{\mathsf{MLE}}\|_{\Sigma_{\mathcal{D}}+\lambda I} \cdot \Big\| \mathop{\mathbb{E}}_{\substack{x\sim\rho,\\ y\sim\pi^\star(\cdot|x)}} [\phi(x,y)] \Big\|_{(\Sigma_{\mathcal{D}}+\lambda I)^{-1}} \\ &\le \mathcal{O}\bigg(\Big((3+\exp(C))\sqrt{\frac{d+\log(1/\delta)}{N}} + \sqrt{\lambda C^2}\Big) \cdot \Big\| \mathop{\mathbb{E}}_{\substack{x\sim\rho,\\ y\sim\pi^\star(\cdot|x)}} [\phi(x,y)] \Big\|_{(\Sigma_{\mathcal{D}}+\lambda I)^{-1}} \bigg).\end{aligned} \tag{G.25}$$

Step 2: bounding term (ib). For the second term of (G.23), recall that

$$\widehat{r} = \arg\min_{r\in\mathcal{R}} \big\{ \ell(r, \mathcal{D}) + \alpha J(r, \widehat{\pi}) \big\},$$

or equivalently

$$\widehat{\theta} = \arg\min_{\theta\in\Theta} \big\{ \ell(r_\theta, \mathcal{D}) + \alpha J(r_\theta, \widehat{\pi}) \big\},$$

and that

$$\theta_{\mathsf{MLE}} = \arg\min_{\theta\in\Theta} \ell(r_\theta, \mathcal{D}).$$

With linear constraint (8.19), by KKT condition we have

$$\nabla_\theta \ell(\widehat{r}, \mathcal{D}) + \alpha \mathop{\mathbb{E}}_{\substack{x\sim\rho,\\ y\sim\widehat{\pi}(\cdot|x)}} [\phi(x,y)] + \lambda_1 \mathop{\mathbb{E}}_{\substack{x\sim\rho,\\ y\sim\pi_{\mathrm{cal}}(\cdot|x)}} [\phi(x,y)] = 0$$

for some $\lambda_1 \in \mathbb{R}$, and

$$\nabla_\theta \ell(r_{\mathsf{MLE}}, \mathcal{D}) + \lambda_2 \mathop{\mathbb{E}}_{\substack{x\sim\rho,\\ y\sim\pi_{\mathrm{cal}}(\cdot|x)}} [\phi(x,y)] = 0$$

for some $\lambda_2 \in \mathbb{R}$. By strong monotonicity of $\nabla_\theta \ell$ (cf. (G.24)), we have

$$\begin{aligned}\frac{N}{3+\exp(C)} \big\|\widehat{\theta} - \theta_{\mathsf{MLE}}\big\|_{\Sigma_{\mathcal{D}}}^2 &\le \Big\langle \nabla_\theta \ell(\widehat{r}, \mathcal{D}) - \nabla_\theta \ell(r_{\mathsf{MLE}}, \mathcal{D}), \widehat{\theta} - \theta_{\mathsf{MLE}} \Big\rangle \\ &= \Big\langle -\alpha \mathop{\mathbb{E}}_{\substack{x\sim\rho,\\ y\sim\widehat{\pi}(\cdot|x)}} [\phi(x,y)] - (\lambda_1 - \lambda_2) \mathop{\mathbb{E}}_{\substack{x\sim\rho,\\ y\sim\pi_{\mathrm{cal}}(\cdot|x)}} [\phi(x,y)], \widehat{\theta} - \theta_{\mathsf{MLE}} \Big\rangle \\ &= -\alpha \Big\langle \mathop{\mathbb{E}}_{\substack{x\sim\rho,\\ y\sim\widehat{\pi}(\cdot|x)}} [\phi(x,y)] - \mathop{\mathbb{E}}_{\substack{x\sim\rho,\\ y\sim\pi_{\mathrm{cal}}(\cdot|x)}} [\phi(x,y)], \widehat{\theta} - \theta_{\mathsf{MLE}} \Big\rangle \\ &\le \alpha \Big\| \mathop{\mathbb{E}}_{\substack{x\sim\rho,\\ y\sim\widehat{\pi}(\cdot|x)}} [\phi(x,y)] - \mathop{\mathbb{E}}_{\substack{x\sim\rho,\\ y\sim\pi_{\mathrm{cal}}(\cdot|x)}} [\phi(x,y)] \Big\|_{(\Sigma_{\mathcal{D}}+\lambda I)^{-1}} \big\|\widehat{\theta} - \theta_{\mathsf{MLE}}\big\|_{\Sigma_{\mathcal{D}}+\lambda I} \\ &\le \alpha\kappa_{\mathcal{D}} \big\|\widehat{\theta} - \theta_{\mathsf{MLE}}\big\|_{\Sigma_{\mathcal{D}}+\lambda I},\end{aligned}$$

where we denote

$$\kappa_{\mathcal{D}} = \Big\| \mathop{\mathbb{E}}_{\substack{x\sim\rho,\\ y\sim\widehat{\pi}(\cdot|x)}} [\phi(x,y)] - \mathop{\mathbb{E}}_{\substack{x\sim\rho,\\ y\sim\pi_{\mathrm{cal}}(\cdot|x)}} [\phi(x,y)] \Big\|_{(\Sigma_{\mathcal{D}}+\lambda I)^{-1}}. \tag{G.26}$$

The penultimate step results from $\widehat{\theta}, \theta_{\mathsf{MLE}} \in \Theta$, which ensures

$$\left\langle \mathop{\mathbb{E}}_{\substack{x\sim\rho,\\ y\sim\pi_{\mathrm{cal}}(\cdot|x)}} [\phi(x,y)], \widehat{\theta} \right\rangle = \left\langle \mathop{\mathbb{E}}_{\substack{x\sim\rho,\\ y\sim\pi_{\mathrm{cal}}(\cdot|x)}} [\phi(x,y)], \theta_{\mathsf{MLE}} \right\rangle = 0$$

It follows that

$$\begin{aligned}
\frac{N}{3+\exp(C)} \big\|\widehat{\theta} - \theta_{\mathsf{MLE}}\big\|^2_{\Sigma_{\mathcal{D}}+\lambda I} &\le \frac{N}{3+\exp(C)} \big\|\widehat{\theta} - \theta_{\mathsf{MLE}}\big\|^2_{\Sigma_{\mathcal{D}}} + \frac{N}{3+\exp(C)} \big\|\widehat{\theta} - \theta_{\mathsf{MLE}}\big\|^2_{\lambda I} \\
&\le \alpha\kappa_{\mathcal{D}} \big\|\widehat{\theta} - \theta_{\mathsf{MLE}}\big\|_{\Sigma_{\mathcal{D}}+\lambda I} + \frac{N\lambda C^2}{3+\exp(C)}.
\end{aligned}$$

The above inequality allows us to bound

$$\big\|\widehat{\theta} - \theta_{\mathsf{MLE}}\big\|_{\Sigma_{\mathcal{D}}+\lambda I} \le \frac{\alpha(3+\exp(C))}{N}\kappa_{\mathcal{D}} + 2\sqrt{\lambda C^2}. \tag{G.27}$$

Therefore, the second term of (G.23) can be bounded as

$$\begin{aligned}
\textbf{Term (ib)} &\le \big\|\widehat{\theta} - \theta_{\mathsf{MLE}}\big\|_{\Sigma_{\mathcal{D}}+\lambda I} \Big\| \mathop{\mathbb{E}}_{\substack{x\sim\rho,\\ y\sim\pi^\star(\cdot|x)}} [\phi(x,y)] \Big\|_{(\Sigma_{\mathcal{D}}+\lambda I)^{-1}} \\
&\le \left(\frac{\alpha(3+\exp(C))}{N}\kappa_{\mathcal{D}} + 2\sqrt{\lambda C^2}\right) \Big\| \mathop{\mathbb{E}}_{\substack{x\sim\rho,\\ y\sim\pi^\star(\cdot|x)}} [\phi(x,y)] \Big\|_{(\Sigma_{\mathcal{D}}+\lambda I)^{-1}}.
\end{aligned} \tag{G.28}$$

Putting (G.25) and (G.28) together, we have

$$\textbf{Term (i)} \le \mathcal{O}\left(\left[\frac{3+\exp(C)}{\sqrt{N}} \left(\sqrt{d + \log(1/\delta)} + \frac{\alpha}{\sqrt{N}}\kappa_{\mathcal{D}} \right) + \sqrt{\lambda C^2} \right] \cdot \Big\| \mathop{\mathbb{E}}_{\substack{x\sim\rho,\\ y\sim\pi^\star(\cdot|x)}} [\phi(x,y)] \Big\|_{(\Sigma_{\mathcal{D}}+\lambda I)^{-1}} \right). \tag{G.29}$$

Step 3: bounding term (ii). We can decompose and bound term (ii) by

$$\begin{aligned}
J(\widehat{r},\widehat{\pi}) - J(r^\star,\widehat{\pi}) &= J(\widehat{r},\widehat{\pi}) + \frac{1}{\alpha}\ell(\widehat{r},\mathcal{D}) - \Big(J(r^\star,\widehat{\pi}) + \frac{1}{\alpha}\ell(r^\star,\mathcal{D}) \Big) + \frac{1}{\alpha}(\ell(\widehat{r},\mathcal{D}) - \ell(r^\star,\mathcal{D})) \\
&\overset{(i)}{\le} \frac{1}{\alpha}(\ell(\widehat{r},\mathcal{D}) - \ell(r^\star,\mathcal{D})) \\
&\le \frac{1}{\alpha}(\ell(\widehat{r},\mathcal{D}) - \ell(r_{\mathsf{MLE}},\mathcal{D}) + \ell(r_{\mathsf{MLE}},\mathcal{D}) - \ell(r^\star,\mathcal{D})),
\end{aligned}$$

where (i) follows from the fact that $(\widehat{r}, \widehat{\pi})$ is a saddle point. Due to convexity of ℓ, we have

$$\begin{aligned}
\ell(\widehat{r}, \mathcal{D}) - \ell(r_{\mathsf{MLE}}, \mathcal{D}) &\leq \left\langle \nabla_\theta \ell(\widehat{r}, \mathcal{D}), \widehat{\theta} - \theta_{\mathsf{MLE}} \right\rangle \\
&= \left\langle -\alpha \mathop{\mathbb{E}}_{\substack{x\sim\rho,\\ y\sim\widehat{\pi}(\cdot|x)}} [\phi(x,y)] - \lambda_1 \mathop{\mathbb{E}}_{\substack{x\sim\rho,\\ y\sim\pi_{\mathrm{cal}}(\cdot|x)}} [\phi(x,y)], \widehat{\theta} - \theta_{\mathsf{MLE}} \right\rangle \\
&= -\alpha \left\langle \mathop{\mathbb{E}}_{\substack{x\sim\rho,\\ y\sim\widehat{\pi}(\cdot|x)}} [\phi(x,y)] - \mathop{\mathbb{E}}_{\substack{x\sim\rho,\\ y\sim\pi_{\mathrm{cal}}(\cdot|x)}} [\phi(x,y)], \widehat{\theta} - \theta_{\mathsf{MLE}} \right\rangle \\
&\leq \alpha \kappa_{\mathcal{D}} \|\widehat{\theta} - \theta_{\mathsf{MLE}}\|_{\Sigma_{\mathcal{D}}+\lambda I} \\
&\leq \frac{\alpha^2(3+\exp(C))}{N} \kappa_{\mathcal{D}}^2 + 2\sqrt{\lambda C^2}\alpha\kappa_{\mathcal{D}},
\end{aligned}$$

where the last step is due to (G.27). On the other hand, with probability $1-\delta$ we have [Zhan et al., 2023b, Lemma 1]:

$$\ell(r_{\mathsf{MLE}}, \mathcal{D}) - \ell(r^\star, \mathcal{D}) \leq \widetilde{\mathcal{O}}(1).$$

Putting pieces together,

$$\textbf{Term (ii)} \leq \frac{\alpha(3+\exp(C))}{N} \kappa_{\mathcal{D}}^2 + 2\sqrt{\lambda C^2}\kappa_{\mathcal{D}} + \frac{1}{\alpha}. \tag{G.30}$$

Step 4: putting things together. Combining (G.22) (G.29), (G.30), with probability $1-\delta$ we have

$$\begin{aligned}
&J^\star(r^\star) - J(r^\star, \widehat{\pi}) \\
&\leq \mathcal{O}\Bigg(\frac{1}{\sqrt{N}} \left[(3+\exp(C)) \left(\sqrt{d + \log(1/\delta)} + \kappa_{\mathcal{D}} \right) + C \right] \cdot \Big\| \mathop{\mathbb{E}}_{\substack{x\sim\rho,\\ y\sim\pi^\star(\cdot|x)}} [\phi(x,y)] \Big\|_{(\Sigma_{\mathcal{D}}+\lambda I)^{-1}} \\
&\quad + \frac{1}{\sqrt{N}} \Big((3+\exp(C))\kappa_{\mathcal{D}}^2 + 2C\kappa_{\mathcal{D}} + 1 \Big) \Bigg).
\end{aligned}$$

Here we have set $\alpha = \sqrt{N}$ and $\lambda = 1/N$. We conclude by bounding $\kappa_{\mathcal{D}}$ as

$$\begin{aligned}
\kappa_{\mathcal{D}}^2 &= \Big\| \mathop{\mathbb{E}}_{\substack{x\sim\rho,\\ y\sim\widehat{\pi}(\cdot|x)}} [\phi(x,y)] - \mathop{\mathbb{E}}_{\substack{x\sim\rho,\\ y\sim\pi_{\mathrm{cal}}(\cdot|x)}} [\phi(x,y)] \Big\|^2_{(\Sigma_{\mathcal{D}}+\lambda I)^{-1}} \\
&\leq \Big\| \mathop{\mathbb{E}}_{\substack{x\sim\rho,\\ y\sim\widehat{\pi}(\cdot|x)}} [\phi(x,y)] - \mathop{\mathbb{E}}_{\substack{x\sim\rho,\\ y\sim\pi_{\mathrm{cal}}(\cdot|x)}} [\phi(x,y)] \Big\|_2^2 \cdot \Big\| (\Sigma_{\mathcal{D}} + \lambda I)^{-1} \Big\|_2 \\
&\leq 4(\lambda_{\min}(\Sigma_{\mathcal{D}}) + \lambda)^{-1}.
\end{aligned}$$

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.

Alekh Agarwal and John C. Duchi. Distributed delayed stochastic optimization. *Advances in neural information processing systems*, 24, 2011.

Alekh Agarwal, Nan Jiang, and Sham M. Kakade. Reinforcement learning: Theory and algorithms. Technical report, 2019.

Alekh Agarwal, Sham Kakade, and Lin F. Yang. Model-based reinforcement learning with a generative model is minimax optimal. In *Conference on Learning Theory*, pages 67–83. PMLR, 2020a.

Alekh Agarwal, Sham M. Kakade, Jason D. Lee, and Gaurav Mahajan. Optimality and approximation with policy gradient methods in Markov decision processes. In *Conference on Learning Theory*, pages 64–66. PMLR, 2020b.

Andrea Agazzi and Jianfeng Lu. Global optimality of softmax policy gradient with single hidden layer neural networks in the mean-field regime, 2020.

Zafarali Ahmed, Nicolas Le Roux, Mohammad Norouzi, and Dale Schuurmans. Understanding the impact of entropy on policy optimization. In *International Conference on Machine Learning*, pages 151–160, 2019.

Ahmet Alacaoglu, Luca Viano, Niao He, and Volkan Cevher. A natural actor-critic framework for zero-sum Markov games. In *International Conference on Machine Learning*, pages 307–366. PMLR, 2022.

Shun-Ichi Amari. Natural gradient works efficiently in learning. *Neural computation*, 10(2):251–276, 1998.

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in AI safety, 2016.

Ioannis Anagnostides, Constantinos Daskalakis, Gabriele Farina, Maxwell Fishelson, Noah Golowich, and Tuomas Sandholm. Near-optimal no-regret learning for correlated equilibria in multi-player general-sum games. In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*, pages 736–749, 2022a.

Ioannis Anagnostides, Ioannis Panageas, Gabriele Farina, and Tuomas Sandholm. On last-iterate convergence beyond zero-sum games. In *International Conference on Machine Learning*, pages 536–581. PMLR, 2022b.

Ruicheng Ao, Shicong Cen, and Yuejie Chi. Asynchronous gradient play in zero-sum multi-agent games. *International Conference on Learning Representations (ICLR 2023)*, 2023.

Sanjeev Arora, Elad Hazan, and Satyen Kale. The multiplicative weights update method: a meta-algorithm and applications. *Theory of Computing*, 8(1):121–164, 2012.

Gürdal Arslan and Serdar Yüksel. Decentralized q-learning for stochastic teams and games. *IEEE Transactions on Automatic Control*, 62(4):1545–1558, 2016.

Mahmoud Assran, Arda Aytekin, Hamid Reza Feyzmahdavian, Mikael Johansson, and Michael G. Rabbat. Advances in asynchronous parallel and distributed optimization. *Proceedings of the IEEE*, 108(11):2013–2031, 2020.

Peter Auer, Thomas Jaksch, and Ronald Ortner. Near-optimal regret bounds for reinforcement learning. *Advances in neural information processing systems*, 21, 2008.

Mohammad Gheshlaghi Azar, Rémi Munos, and Hilbert J. Kappen. Minimax PAC bounds on the sample complexity of reinforcement learning with a generative model. *Machine learning*, 91(3): 325–349, 2013.

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 263–272. JMLR. org, 2017.

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.

Yu Bai and Chi Jin. Provable self-play algorithms for competitive reinforcement learning. In *International Conference on Machine Learning*, pages 551–560. PMLR, 2020.

Yu Bai, Tengyang Xie, Nan Jiang, and Yu-Xiang Wang. Provably efficient Q-learning with low switching cost. *Advances in Neural Information Processing Systems*, 32, 2019.

Yu Bai, Chi Jin, and Tiancheng Yu. Near-optimal reinforcement learning with self-play. *Advances in neural information processing systems*, 33:2159–2170, 2020.

James P. Bailey. $O(1/T)$ time-average convergence in a generalization of multiagent zero-sum games. *arXiv preprint arXiv:2110.02482*, 2021.

Amir Beck. *First-order methods in optimization*. SIAM, 2017.

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

Richard Bellman. On the theory of dynamic programming. *Proceedings of the National Academy of Sciences of the United States of America*, 38(8):716, 1952.

L.M. Bergman and I.N. Fokin. On separable non-cooperative zero-sum games. *Optimization*, 44 (1):69–84, 1998.

Riccardo Bernardi. Interactive image segmentation using graph transduction games. 2021.

Dimitri P. Bertsekas. *Dynamic programming and optimal control (4th edition)*. Athena Scientific, 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*, 2024.

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

Michael Bowling and Manuela Veloso. Rational and convergent learning in stochastic games. In *Proceedings of the 17th international joint conference on Artificial intelligence-Volume 2*, pages 1021–1026, 2001.

Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

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.

Yang Cai, Ozan Candogan, Constantinos Daskalakis, and Christos Papadimitriou. Zero-sum polymatrix games: A generalization of minmax. *Mathematics of Operations Research*, 41(2):648–655, 2016.

Shicong Cen, Yuting Wei, and Yuejie Chi. Fast policy extragradient methods for competitive games with entropy regularization. *Advances in Neural Information Processing Systems*, 34: 27952–27964, 2021.

Shicong Cen, Fan Chen, and Yuejie Chi. Independent natural policy gradient methods for potential games: Finite-time global convergence with entropy regularization. In *2022 IEEE 61th Conference on Decision and Control (CDC)*. IEEE, 2022a.

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, 2022b.

Shicong Cen, Yuejie Chi, Simon Du, and Lin Xiao. Faster last-iterate convergence of policy optimization in zero-sum Markov games. *International Conference on Learning Representations (ICLR 2023)*, 2023.

Shicong Cen, Jincheng Mei, Katayoon Goshvadi, Hanjun Dai, Tong Yang, Sherry Yang, Dale Schuurmans, Yuejie Chi, and Bo Dai. Value-incentivized preference optimization: A unified approach to online and offline RLHF. *arXiv preprint arXiv:2405.19320*, 2024.

Nicolò Cesa-Bianchi, Claudio Gentile, Yishay Mansour, and Alberto Minora. Delay and cooperation in nonstochastic bandits. In *Conference on Learning Theory*, pages 605–622. PMLR, 2016.

Nicolò Cesa-Bianchi, Claudio Gentile, Gábor Lugosi, and Gergely Neu. Boltzmann exploration done right. *Advances in neural information processing systems*, 30, 2017.

Jonathan D. Chang, Wenhao Shan, Owen Oertell, Kianté Brantley, Dipendra Misra, Jason D. Lee, and Wen Sun. Dataset reset policy optimization for RLHF. *arXiv preprint arXiv:2404.08495*, 2024.

Zixiang Chen, Dongruo Zhou, and Quanquan Gu. Almost optimal algorithms for two-player zero-sum linear mixture Markov games. In *International Conference on Algorithmic Learning Theory*, pages 227–261. PMLR, 2022.

Zixiang Chen, Yihe Deng, Huizhuo Yuan, Kaixuan Ji, and Quanquan Gu. Self-play fine-tuning converts weak language models to strong language models. *arXiv preprint arXiv:2401.01335*, 2024.

Ziyi Chen, Shaocong Ma, and Yi Zhou. Sample efficient stochastic policy extragradient algorithm for zero-sum Markov game. In *International Conference on Learning Representations*, 2021.

Steve Chien and Alistair Sinclair. Convergence to approximate nash equilibria in congestion games. *Games and Economic Behavior*, 71(2):315–327, 2011.

Yinlam Chow, Ofir Nachum, Edgar Duenez-Guzman, and Mohammad Ghavamzadeh. A Lyapunov-based approach to safe reinforcement learning. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 8103–8112, 2018a.

Yinlam Chow, Ofir Nachum, and Mohammad Ghavamzadeh. Path consistency learning in Tsallis entropy regularized MDPs. In *International Conference on Machine Learning*, pages 979–988. PMLR, 2018b.

George Christodoulou, Vahab S. Mirrokni, and Anastasios Sidiropoulos. Convergence and approximation in potential games. In *Annual Symposium on Theoretical Aspects of Computer Science*, pages 349–360. Springer, 2006.

Bo Dai, Albert Shaw, Lihong Li, Lin Xiao, Niao He, Zhen Liu, Jianshu Chen, and Le Song. SBEED: Convergent reinforcement learning with nonlinear function approximation. In *International Conference on Machine Learning*, pages 1125–1134. PMLR, 2018.

Constantinos Daskalakis. On the complexity of approximating a Nash equilibrium. *ACM Transactions on Algorithms (TALG)*, 9(3):1–35, 2013.

Constantinos Daskalakis and Ioannis Panageas. The limit points of (optimistic) gradient descent in min-max optimization. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 9256–9266, 2018.

Constantinos Daskalakis and Ioannis Panageas. Last-iterate convergence: Zero-sum games and constrained min-max optimization. In *10th Innovations in Theoretical Computer Science Conference (ITCS 2019)*. Schloss-Dagstuhl-Leibniz Zentrum für Informatik, 2019.

Constantinos Daskalakis, Alan Deckelbaum, and Anthony Kim. Near-optimal no-regret algorithms for zero-sum games. In *Proceedings of the twenty-second annual ACM-SIAM symposium on Discrete Algorithms*, pages 235–254. SIAM, 2011.

Constantinos Daskalakis, Andrew Ilyas, Vasilis Syrgkanis, and Haoyang Zeng. Training GANs with optimism. In *International Conference on Learning Representations (ICLR 2018)*, 2018.

Constantinos Daskalakis, Dylan J. Foster, and Noah Golowich. Independent policy gradient methods for competitive reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 5527–5540, 2020.

Constantinos Daskalakis, Maxwell Fishelson, and Noah Golowich. Near-optimal no-regret learning in general games. *Advances in Neural Information Processing Systems*, 34, 2021.

Dongsheng Ding, Xiaohan Wei, Zhuoran Yang, Zhaoran Wang, and Mihailo Jovanovic. Provably efficient safe exploration via primal-dual policy optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 3304–3312. PMLR, 2021.

Dongsheng Ding, Chen-Yu Wei, Kaiqing Zhang, and Mihailo R. Jovanović. Independent policy gradient for large-scale Markov potential games: Sharper rates, function approximation, and game-agnostic convergence. In *International Conference on Machine Learning*, pages 5166–5220. PMLR, 2022.

Bruce K. Driver. Math 280 (probability theory) lecture notes, April 2007. URL: https://mathweb.ucsd.edu/~bdriver/280_06-07/Lecture_Notes/N18_2p.pdf.

Yan Duan, Xi Chen, Rein Houthooft, John Schulman, and Pieter Abbeel. Benchmarking deep reinforcement learning for continuous control. In *International Conference on Machine Learning*, pages 1329–1338, 2016.

John C. Duchi, Shai Shalev-Shwartz, Yoram Singer, and Ambuj Tewari. Composite objective mirror descent. In *Proceedings of the Twenty Third Annual Conference on Computational Learning Theory*, 2010.

Yonathan Efroni, Shie Mannor, and Matteo Pirotta. Exploration-exploitation in constrained MDPs. *arXiv preprint arXiv:2003.02189*, 2020.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. KTO: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*, 2024.

Eyal Even-Dar, Sham M. Kakade, and Yishay Mansour. Online Markov decision processes. *Mathematics of Operations Research*, 34(3):726–736, 2009.

Maryam Fazel, Rong Ge, Sham Kakade, and Mehran Mesbahi. Global convergence of policy gradient methods for the linear quadratic regulator. In *International Conference on Machine Learning*, pages 1467–1476, 2018.

Genevieve E. Flaspohler, Francesco Orabona, Judah Cohen, Soukayna Mouatadid, Miruna Oprescu, Paulo Orenstein, and Lester Mackey. Online learning with optimism and delay. In *International Conference on Machine Learning*, pages 3363–3373. PMLR, 2021.

Roy Fox, Stephen M. Mcaleer, Will Overman, and Ioannis Panageas. Independent natural policy gradient always converges in Markov potential games. In *International Conference on Artificial Intelligence and Statistics*, pages 4414–4425. PMLR, 2022.

Yoav Freund and Robert E. Schapire. Adaptive game playing using multiplicative weights. *Games and Economic Behavior*, 29(1-2):79–103, 1999.

Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, et al. A survey of uncertainty in deep neural networks. *Artificial Intelligence Review*, 56(Suppl 1):1513–1589, 2023.

Matthieu Geist, Bruno Scherrer, and Olivier Pietquin. A theory of regularized Markov decision processes. In *International Conference on Machine Learning*, pages 2160–2169, 2019.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

Jean-Bastien Grill, Omar Darwiche Domingues, Pierre Menard, Remi Munos, and Michal Valko. Planning in entropy-regularized Markov decision processes and games. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Shangmin Guo, Biao Zhang, Tianlin Liu, Tianqi Liu, Misha Khalman, Felipe Llinares, Alexandre Rame, Thomas Mesnard, Yao Zhao, Bilal Piot, et al. Direct language model alignment from online AI feedback. *arXiv preprint arXiv:2402.04792*, 2024.

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, 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, 2018.

Botao Hao, Nevena Lazic, Yasin Abbasi-Yadkori, Pooria Joulani, and Csaba Szepesvári. Adaptive approximate policy iteration. In *International Conference on Artificial Intelligence and Statistics*, pages 523–531. PMLR, 2021.

Patrick T. Harker and Jong-Shi Pang. Finite-dimensional variational inequality and nonlinear complementarity problems: a survey of theory, algorithms and applications. *Mathematical programming*, 48(1):161–220, 1990.

Elad Hazan, Sham Kakade, Karan Singh, and Abby Van Soest. Provably efficient maximum entropy exploration. In *International Conference on Machine Learning*, pages 2681–2691, 2019.

Xinran He, Junfeng Pan, Ou Jin, Tianbing Xu, Bo Liu, Tao Xu, Yanxin Shi, Antoine Atallah, Ralf Herbrich, Stuart Bowers, et al. Practical lessons from predicting clicks on ads at facebook. In *Proceedings of the eighth international workshop on data mining for online advertising*, pages 1–9, 2014.

Amélie Heliou, Johanne Cohen, and Panayotis Mertikopoulos. Learning with bandit feedback in potential games. *Advances in Neural Information Processing Systems*, 30, 2017.

Josef Hofbauer and William H. Sandholm. On the global convergence of stochastic fictitious play. *Econometrica*, 70(6):2265–2294, 2002.

Yu-Guan Hsieh, Franck Iutzeler, Jérôme Malick, and Panayotis Mertikopoulos. On the convergence of single-call stochastic extra-gradient methods. *Advances in Neural Information Processing Systems*, 32, 2019.

Yu-Guan Hsieh, Franck Iutzeler, Jérôme Malick, and Panayotis Mertikopoulos. Multi-agent online optimization with delays: Asynchronicity, adaptivity, and optimism. *Journal of Machine Learning Research*, 23(78):1–49, 2022.

Bin Hu, Kaiqing Zhang, Na Li, Mehran Mesbahi, Maryam Fazel, and Tamer Başar. Toward a theoretical foundation of policy optimization for learning control policies. *Annual Review of Control, Robotics, and Autonomous Systems*, 6(1):123–158, 2023.

Yu-Heng Hung, Ping-Chun Hsieh, Xi Liu, and P.R. Kumar. Reward-biased maximum likelihood estimation for linear stochastic bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 7874–7882, 2021.

Joao Paulo Jansch-Porto, Bin Hu, and Geir E. Dullerud. Convergence guarantees of policy optimization methods for Markovian jump linear systems. In *2020 American Control Conference (ACC)*, pages 2882–2887. IEEE, 2020.

Harold Jeffreys. *The theory of probability*. OUP Oxford, 1998.

Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E. Schapire. Contextual decision processes with low Bellman rank are pac-learnable. In *International Conference on Machine Learning*, pages 1704–1713. PMLR, 2017.

Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I. Jordan. Is Q-learning provably efficient? In *Advances in Neural Information Processing Systems*, pages 4863–4873, 2018.

Chi Jin, Qinghua Liu, and Tiancheng Yu. The power of exploiter: Provable multi-agent RL in large state spaces. In *International Conference on Machine Learning*, pages 10251–10279. PMLR, 2022.

Chi Jin, Qinghua Liu, Yuanhao Wang, and Tiancheng Yu. V-learning–a simple, efficient, decentralized algorithm for multiagent RL. *Mathematics of Operations Research*, 2023.

Pooria Joulani, Andras Gyorgy, and Csaba Szepesvári. Online learning under delayed feedback. In *International Conference on Machine Learning*, pages 1453–1461. PMLR, 2013.

Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *Proceedings of the Nineteenth International Conference on Machine Learning*, pages 267–274, 2002.

Sham M. Kakade. A natural policy gradient. In *Advances in neural information processing systems*, pages 1531–1538, 2001.

Ehsan Asadi Kangarshahi, Ya-Ping Hsieh, Mehmet Fatih Sahin, and Volkan Cevher. Let's be honest: An optimal no-regret framework for zero-sum games. In *International Conference on Machine Learning*, pages 2488–2496. PMLR, 2018.

Belhal Karimi, Blazej Miasojedow, Éric Moulines, and Hoi-To Wai. Non-asymptotic analysis of biased stochastic approximation scheme. In *Conference on Learning Theory*, pages 1944–1974. PMLR, 2019.

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.

Bahare Kiumarsi, Kyriakos G. Vamvoudakis, Hamidreza Modares, and Frank L. Lewis. Optimal and autonomous control using reinforcement learning: A survey. *IEEE transactions on neural networks and learning systems*, 29(6):2042–2062, 2017.

Krzysztof C. Kiwiel. Proximal minimization methods with generalized Bregman functions. *SIAM journal on control and optimization*, 35(4):1142–1168, 1997.

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

Galina M. Korpelevich. The extragradient method for finding saddle points and other problems. *Matecon*, 12:747–756, 1976.

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative Q-learning for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 33:1179–1191, 2020.

P. Kumar and A. Becker. A new family of optimal adaptive controllers for Markov chains. *IEEE Transactions on Automatic Control*, 27(1):137–146, 1982.

Tze Leung Lai, Herbert Robbins, et al. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

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

Guanghui Lan and Yi Zhou. An optimal randomized incremental gradient method. *Mathematical programming*, 171(1):167–215, 2018.

Guanghui Lan, Zhaosong Lu, and Renato D.C. Monteiro. Primal-dual first-order methods with $O(1/\varepsilon)$ iteration-complexity for cone programming. *Mathematical Programming*, 126(1):1–29, 2011.

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

Nevena Lazic, Dong Yin, Yasin Abbasi-Yadkori, and Csaba Szepesvari. Improved regret bound and experience replay in regularized policy iteration. In *International Conference on Machine Learning*, pages 6032–6042. PMLR, 2021.

Kyungjae Lee, Sungjoon Choi, and Songhwai Oh. Sparse Markov decision processes with causal sparse Tsallis entropy regularization for reinforcement learning. *IEEE Robotics and Automation Letters*, 3(3):1466–1473, 2018.

Qi Lei, Sai Ganesh Nagarajan, Ioannis Panageas, and Xiao Wang. Last iterate convergence in no-regret learning: constrained min-max optimization for convex-concave landscapes. In *International Conference on Artificial Intelligence and Statistics*, pages 1441–1449. PMLR, 2021.

Stefanos Leonardos, Georgios Piliouras, and Kelly Spendlove. Exploration-exploitation in multi-agent competition: convergence with bounded rationality. *Advances in Neural Information Processing Systems*, 34:26318–26331, 2021.

Stefanos Leonardos, Will Overman, Ioannis Panageas, and Georgios Piliouras. Global convergence of multi-agent policy gradient in Markov potential games. *International Conference on Learning Representations (ICLR 2022)*, 2022.

Bingcong Li, Tianyi Chen, and Georgios B. Giannakis. Bandit online learning with unknown delays. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 993–1002. PMLR, 2019.

Gen Li, Yuting Wei, Yuejie Chi, Yuantao Gu, and Yuxin Chen. Sample complexity of asynchronous Q-learning: Sharper analysis and variance reduction. *IEEE Transactions on Information Theory*, 68(1):448–473, 2021.

Gen Li, Yuejie Chi, Yuting Wei, and Yuxin Chen. Minimax-optimal multi-agent RL in Markov games with a generative model. *Advances in Neural Information Processing Systems*, 35:15353–15367, 2022.

Gen Li, Yuting Wei, Yuejie Chi, and Yuxin Chen. Softmax policy gradient methods can take exponential time to converge. *Mathematical Programming*, 201(1-2):707–802, 2023.

Gen Li, Changxiao Cai, Yuxin Chen, Yuting Wei, and Yuejie Chi. Is Q-learning minimax optimal? a tight sample complexity analysis. *Operations Research*, 72(1):222–236, 2024a.

Gen Li, Yuting Wei, Yuejie Chi, and Yuxin Chen. Breaking the sample size barrier in model-based reinforcement learning with a generative model. *Operations Research*, 72(1):203–221, 2024b.

Tengyuan Liang and James Stokes. Interaction matters: A note on non-asymptotic local convergence of generative adversarial networks. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 907–915. PMLR, 2019.

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. *International Conference on Learning Representations (ICLR 2016)*, 2016.

Nick Littlestone and Manfred K. Warmuth. The weighted majority algorithm. *Information and computation*, 108(2):212–261, 1994.

Michael L. Littman. Markov games as a framework for multi-agent reinforcement learning. In *Machine Learning Proceedings*, pages 157–163. Elsevier, 1994.

Boyi Liu, Qi Cai, Zhuoran Yang, and Zhaoran Wang. Neural trust region/proximal policy optimization attains globally optimal policy. In *Advances in Neural Information Processing Systems*, pages 10565–10576, 2019a.

Ji Liu, Steve Wright, Christopher Ré, Victor Bittorf, and Srikrishna Sridhar. An asynchronous parallel stochastic coordinate descent algorithm. In *International Conference on Machine Learning*, pages 469–477. PMLR, 2014.

Qinghua Liu, Tiancheng Yu, Yu Bai, and Chi Jin. A sharp analysis of model-based reinforcement learning with self-play. In *International Conference on Machine Learning*, pages 7001–7010. PMLR, 2021.

Siqi Liu, Kee Yuan Ngiam, and Mengling Feng. Deep reinforcement learning for clinical decision support: a brief survey. *arXiv preprint arXiv:1907.09475*, 2019b.

Xi Liu, Ping-Chun Hsieh, Yu Heng Hung, Anirban Bhattacharya, and P. Kumar. Exploration through reward biasing: Reward-biased maximum likelihood estimation for stochastic multi-armed bandits. In *International Conference on Machine Learning*, pages 6248–6258. PMLR, 2020a.

Yanli Liu, Kaiqing Zhang, Tamer Basar, and Wotao Yin. An improved analysis of (variance-reduced) policy gradient and natural policy gradient methods. *Advances in Neural Information Processing Systems*, 33, 2020b.

Zhihan Liu, Miao Lu, Wei Xiong, Han Zhong, Hao Hu, Shenao Zhang, Sirui Zheng, Zhuoran Yang, and Zhaoran Wang. Maximize to explore: One objective function fusing estimation, planning, and exploration. *Advances in Neural Information Processing Systems*, 36, 2024a.

Zhihan Liu, Miao Lu, Shenao Zhang, Boyi Liu, Hongyi Guo, Yingxiang Yang, Jose Blanchet, and Zhaoran Wang. Provably mitigating overoptimization in RLHF: Your SFT loss is implicitly an adversarial regularizer. *arXiv preprint arXiv:2405.16436*, 2024b.

Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, Abbeel Pieter, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. *Advances in neural information processing systems*, 30, 2017.

Weichao Mao and Tamer Başar. Provably efficient reinforcement learning in decentralized general-sum Markov games. *Dynamic Games and Applications*, pages 1–22, 2022.

Weichao Mao, Lin Yang, Kaiqing Zhang, and Tamer Başar. On improving model-free algorithms for decentralized multi-agent reinforcement learning. In *International Conference on Machine Learning*, pages 15007–15049. PMLR, 2022.

Jason R. Marden, Gürdal Arslan, and Jeff S. Shamma. Regret based dynamics: convergence in weakly acyclic games. In *Proceedings of the 6th international joint conference on Autonomous agents and multiagent systems*, pages 1–8, 2007.

Jason R. Marden, H. Peyton Young, Gürdal Arslan, and Jeff S. Shamma. Payoff-based dynamics for multiplayer weakly acyclic games. *SIAM Journal on Control and Optimization*, 48(1):373–396, 2009.

Richard D. McKelvey and Thomas R. Palfrey. Quantal response equilibria for normal form games. *Games and economic behavior*, 10(1):6–38, 1995.

Brendan McMahan and Matthew Streeter. Delay-tolerant algorithms for asynchronous distributed online learning. *Advances in Neural Information Processing Systems*, 27, 2014.

Jincheng Mei, Chenjun Xiao, Bo Dai, Lihong Li, Csaba Szepesvári, and Dale Schuurmans. Escaping the gravitational pull of softmax. *Advances in Neural Information Processing Systems*, 33, 2020a.

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, 2020b.

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, 2021.

Min Meng, Xiuxian Li, and Jie Chen. Decentralized nash equilibria learning for online game with bandit feedback. *IEEE Transactions on Automatic Control*, 2023.

Panayotis Mertikopoulos and William H. Sandholm. Learning in games via reinforcement and regularization. *Mathematics of Operations Research*, 41(4):1297–1324, 2016.

Panayotis Mertikopoulos, Bruno Lecouat, Houssam Zenati, Chuan-Sheng Foo, Vijay Chandrasekhar, and Georgios Piliouras. Optimistic mirror descent in saddle-point problems: Going the extra (gradient) mile. In *International Conference on Learning Representations*, 2018a.

Panayotis Mertikopoulos, Christos Papadimitriou, and Georgios Piliouras. Cycles in adversarial regularized learning. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2703–2717. SIAM, 2018b.

Akshay Mete, Rahul Singh, Xi Liu, and P.R. Kumar. Reward biased maximum likelihood estimation for reinforcement learning. In *Learning for Dynamics and Control*, pages 815–827. PMLR, 2021.

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.

Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*, pages 1928–1937, 2016.

Hesameddin Mohammadi, Armin Zare, Mahdi Soltanolkotabi, and Mihailo R. Jovanović. Convergence and sample complexity of gradient methods for the model-free linear–quadratic regulator problem. *IEEE Transactions on Automatic Control*, 67(5):2435–2450, 2021.

Aryan Mokhtari, Asuman Ozdaglar, and Sarath Pattathil. A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: Proximal point approach. In *International Conference on Artificial Intelligence and Statistics*, pages 1497–1507. PMLR, 2020a.

Aryan Mokhtari, Asuman E. Ozdaglar, and Sarath Pattathil. Convergence rate of $O(1/k)$ for optimistic gradient and extragradient methods in smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 30(4):3230–3251, 2020b.

Teodor Mihai Moldovan and Pieter Abbeel. Safe exploration in Markov decision processes, 2012.

Dov Monderer and Lloyd S. Shapley. Fictitious play property for games with identical interests. *Journal of economic theory*, 68(1):258–265, 1996a.

Dov Monderer and Lloyd S. Shapley. Potential games. *Games and economic behavior*, 14(1): 124–143, 1996b.

Rémi Munos, Michal Valko, Daniele Calandriello, Mohammad Gheshlaghi Azar, Mark Rowland, Zhaohan Daniel Guo, Yunhao Tang, Matthieu Geist, Thomas Mesnard, Andrea Michi, et al. Nash learning from human feedback. *arXiv preprint arXiv:2312.00886*, 2023.

Ofir Nachum, Mohammad Norouzi, Kelvin Xu, and Dale Schuurmans. Bridging the gap between value and policy based reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 2775–2785, 2017.

John Nash. Non-cooperative games. *Annals of mathematics*, pages 286–295, 1951.

Arkadi Nemirovski. Prox-method with rate of convergence $O(1/t)$ for variational inequalities with Lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15(1):229–251, 2004.

Arkadi Semenovich Nemirovsky and David Borisovich Yudin. Problem complexity and method efficiency in optimization. 1983.

Yurii Nesterov. Primal-dual subgradient methods for convex problems. *Mathematical programming*, 120(1):221–259, 2009.

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

J. V. Neumann. Zur theorie der gesellschaftsspiele. *Mathematische annalen*, 100(1):295–320, 1928.

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.

Arka Pal, Deep Karkhanis, Samuel Dooley, Manley Roberts, Siddartha Naidu, and Colin White. Smaug: Fixing failure modes of preference optimisation with DPO-Positive. *arXiv preprint arXiv:2402.13228*, 2024.

Gerasimos Palaiopanos, Ioannis Panageas, and Georgios Piliouras. Multiplicative weights update with constant step-size in congestion games: Convergence, limit cycles and chaos. *Advances in Neural Information Processing Systems*, 30, 2017.

Richard Yuanzhe Pang, Weizhe Yuan, Kyunghyun Cho, He He, Sainbayar Sukhbaatar, and Jason Weston. Iterative reasoning preference optimization. *arXiv preprint arXiv:2404.19733*, 2024.

Stephen D. Patek and Dimitri P. Bertsekas. Stochastic shortest path games. *SIAM Journal on Control and Optimization*, 37(3):804–824, 1999.

Sarath Pattathil, Kaiqing Zhang, and Asuman Ozdaglar. Symmetric (optimistic) natural policy gradient for multi-agent learning with parameter convergence. In *International Conference on Artificial Intelligence and Statistics*, pages 5641–5685. PMLR, 2023.

Julien Perolat, Bruno Scherrer, Bilal Piot, and Olivier Pietquin. Approximate dynamic programming for two-player zero-sum Markov games. In *International Conference on Machine Learning*, pages 1321–1329. PMLR, 2015.

Jan Peters and Stefan Schaal. Natural actor-critic. *Neurocomputing*, 71(7-9):1180–1190, 2008.

Jan Peters, Katharina Mulling, and Yasemin Altun. Relative entropy policy search. In *Twenty-Fourth AAAI Conference on Artificial Intelligence*, 2010.

Ciara Pike-Burke, Shipra Agrawal, Csaba Szepesvari, and Steffen Grunewalder. Bandits with delayed, aggregated anonymous feedback. In *International Conference on Machine Learning*, pages 4105–4113. PMLR, 2018.

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

Kent Quanrud and Daniel Khashabi. Online learning with adversarial delays. *Advances in neural information processing systems*, 28, 2015.

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, 2023.

Sasha Rakhlin and Karthik Sridharan. Optimization, learning, and games with predictable sequences. *Advances in Neural Information Processing Systems*, 26, 2013.

Paria Rashidinejad, Banghua Zhu, Cong Ma, Jiantao Jiao, and Stuart Russell. Bridging offline reinforcement learning and imitation learning: A tale of pessimism. *IEEE Transactions on Information Theory*, 68(12):8156–8196, 2022.

Benjamin Recht, Christopher Re, Stephen Wright, and Feng Niu. Hogwild!: A lock-free approach to parallelizing stochastic gradient descent. *Advances in neural information processing systems*, 24, 2011.

Yagiz Savas, Mohamadreza Ahmadi, Takashi Tanaka, and Ufuk Topcu. Entropy-regularized stochastic games. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pages 5955–5962. IEEE, 2019.

Muhammed Sayin, Kaiqing Zhang, David Leslie, Tamer Başar, and Asuman Ozdaglar. Decentralized Q-learning in zero-sum Markov games. *Advances in Neural Information Processing Systems*, 34:18320–18334, 2021.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International conference on machine learning*, pages 1889–1897, 2015.

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.

Reinhard Selten. Evolution, learning and economic behavior. 1989.

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.

Lloyd S. Shapley. Stochastic games. *Proceedings of the National Academy of Sciences*, 39(10): 1095–1100, 1953.

David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of Go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

Samuel Sokota, Ryan D'Orazio, J. Zico Kolter, Nicolas Loizou, Marc Lanctot, Ioannis Mitliagkas, Noam Brown, and Christian Kroer. A unified approach to reinforcement learning, quantal response equilibria, and two-player zero-sum games. *International Conference on Learning Representations (ICLR 2023)*, 2023.

Ziang Song, Song Mei, and Yu Bai. When can we learn general-sum Markov games with a large number of players sample-efficiently? *International Conference on Learning Representations (ICLR 2022)*, 2022.

Richard S. Sutton and Andrew G. Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

Richard S. Sutton, David A. McAllester, Satinder P. Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in neural information processing systems*, pages 1057–1063, 2000.

Gokul Swamy, Christoph Dann, Rahul Kidambi, Zhiwei Steven Wu, and Alekh Agarwal. A minimaximalist approach to reinforcement learning from human feedback. *arXiv preprint arXiv:2401.04056*, 2024.

Vasilis Syrgkanis, Alekh Agarwal, Haipeng Luo, and Robert E. Schapire. Fast convergence of regularized learning in games. In *Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 2*, pages 2989–2997, 2015.

Yunhao Tang, Zhaohan Daniel Guo, Zeyu Zheng, Daniele Calandriello, Rémi Munos, Mark Rowland, Pierre Harvey Richemond, Michal Valko, Bernardo Ávila Pires, and Bilal Piot. Generalized preference optimization: A unified approach to offline alignment. *arXiv preprint arXiv:2402.05749*, 2024.

Manan Tomar, Lior Shani, Yonathan Efroni, and Mohammad Ghavamzadeh. Mirror descent policy optimization. *International Conference on Learning Representations (ICLR 2022)*, 2022.

Paul Tseng. On linear convergence of iterative methods for the variational inequality problem. *Journal of Computational and Applied Mathematics*, 60(1-2):237–252, 1995.

Stephen Tu and Benjamin Recht. The gap between model-based and model-free methods on the linear quadratic regulator: An asymptotic viewpoint. In *Conference on Learning Theory*, pages 3036–3083, 2019.

Masatoshi Uehara and Wen Sun. Pessimistic model-based offline reinforcement learning under partial coverage. *International Conference on Learning Representations (ICLR 2022)*, 2022.

J. Van Der Wal. Discounted Markov games: Generalized policy iteration method. *Journal of Optimization Theory and Applications*, 25(1):125–138, 1978.

Claire Vernade, Olivier Cappé, and Vianney Perchet. Stochastic bandit models for delayed conversions. In *Conference on Uncertainty in Artificial Intelligence*, 2017.

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, 2020.

Yuanyu Wan, Wei-Wei Tu, and Lijun Zhang. Online strongly convex optimization with unknown delays. *Machine Learning*, 111(3):871–893, 2022.

Lingxiao Wang, Qi Cai, Zhuoran Yang, and Zhaoran Wang. Neural policy gradient methods: Global optimality and rates of convergence. *International Conference on Learning Representations (ICLR 2020)*, 2020.

Weichen Wang, Jiequn Han, Zhuoran Yang, and Zhaoran Wang. Global convergence of policy gradient for linear-quadratic mean-field control/game in continuous time. In *International Conference on Machine Learning*, pages 10772–10782. PMLR, 2021.

Chen-Yu Wei, Chung-Wei Lee, Mengxiao Zhang, and Haipeng Luo. Linear last-iterate convergence in constrained saddle-point optimization. In *International Conference on Learning Representations (ICLR)*, 2021a.

Chen-Yu Wei, Chung-Wei Lee, Mengxiao Zhang, and Haipeng Luo. Last-iterate convergence of decentralized optimistic gradient descent/ascent in infinite-horizon competitive Markov games. In *Conference on learning theory*, pages 4259–4299. PMLR, 2021b.

Marcelo J. Weinberger and Erik Ordentlich. On delayed prediction of individual sequences. *IEEE Transactions on Information Theory*, 48(7):1959–1976, 2002.

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

Ronald J. Williams and Jing Peng. Function optimization using connectionist reinforcement learning algorithms. *Connection Science*, 3(3):241–268, 1991.

Yue Wu, Weitong Zhang, Pan Xu, and Quanquan Gu. A finite time analysis of two time-scale actor critic methods, 2020.

Chenjun Xiao, Ruitong Huang, Jincheng Mei, Dale Schuurmans, and Martin Müller. Maximum entropy Monte-Carlo planning. In *Advances in Neural Information Processing Systems*, pages 9520–9528, 2019.

Chenjun Xiao, Yifan Wu, Jincheng Mei, Bo Dai, Tor Lattimore, Lihong Li, Csaba Szepesvari, and Dale Schuurmans. On the optimality of batch policy optimization algorithms. In *International Conference on Machine Learning*, pages 11362–11371. PMLR, 2021.

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

Qiaomin Xie, Yudong Chen, Zhaoran Wang, and Zhuoran Yang. Learning zero-sum simultaneous-move Markov games using function approximation and correlated equilibrium. In *Conference on Learning Theory*, pages 3674–3682. PMLR, 2020.

Tengyang Xie, Dylan J. Foster, Akshay Krishnamurthy, Corby Rosset, Ahmed Awadallah, and Alexander Rakhlin. Exploratory preference optimization: Harnessing implicit Q*-approximation for sample-efficient RLHF. *arXiv preprint arXiv:2405.21046*, 2024.

Wei Xiong, Hanze Dong, Chenlu Ye, Han Zhong, Nan Jiang, and Tong Zhang. Gibbs sampling from human feedback: A provable KL-constrained framework for RLHF. *arXiv preprint arXiv:2312.11456*, 2023.

Pan Xu, Felicia Gao, and Quanquan Gu. Sample efficient policy gradient methods with recursive variance reduction. *International Conference on Learning Representations (ICLR 2020)*, 2020a.

Tengyu Xu, Yingbin Liang, and Guanghui Lan. A primal approach to constrained policy optimization: Global optimality and finite-time analysis. *arXiv preprint arXiv:2011.05869*, 2020b.

Tengyu Xu, Zhe Wang, and Yingbin Liang. Non-asymptotic convergence analysis of two time-scale (natural) actor-critic algorithms. *arXiv preprint arXiv:2005.03557*, 2020c.

Abhay Yadav, Sohil Shah, Zheng Xu, David Jacobs, and Tom Goldstein. Stabilizing adversarial nets with prediction methods. *International Conference on Learning Representations (ICLR 2018)*, 2018.

Tong Yang, Shicong Cen, Yuting Wei, Yuxin Chen, and Yuejie Chi. Federated natural policy gradient methods for multi-task reinforcement learning. *arXiv preprint arXiv:2311.00201*, 2023.

Wenhao Yang, Xiang Li, Guangzeng Xie, and Zhihua Zhang. Finding the near optimal policy via adaptive reduced regularization in MDPs. *arXiv preprint arXiv:2011.00213*, 2020.

Yuepeng Yang and Cong Ma. $\mathcal{O}(T^{-1})$ convergence of optimistic-follow-the-regularized-leader in two-player zero-sum Markov games. *International Conference on Learning Representations (ICLR 2023)*, 2023.

H. Peyton Young. *Strategic learning and its limits*. OUP Oxford, 2004.

H. Peyton Young. Individual strategy and social structure. In *Individual Strategy and Social Structure*. Princeton University Press, 2020.

Ming Yu, Zhuoran Yang, Mladen Kolar, and Zhaoran Wang. Convergent policy optimization for safe reinforcement learning. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 3127–3139, 2019.

Sihan Zeng, Thinh T. Doan, and Justin Romberg. Regularized gradient descent ascent for two-player zero-sum Markov games. *Advances in Neural Information Processing Systems*, 35, 2022.

Wenhao Zhan, Shicong Cen, Baihe Huang, Yuxin Chen, Jason D. Lee, and Yuejie Chi. Policy mirror descent for regularized reinforcement learning: A generalized framework with linear convergence. *SIAM Journal on Optimization*, 33(2):1061–1091, 2023a.

Wenhao Zhan, Masatoshi Uehara, Nathan Kallus, Jason D. Lee, and Wen Sun. Provable offline preference-based reinforcement learning. In *The Twelfth International Conference on Learning Representations*, 2023b.

Junyu Zhang, Alec Koppel, Amrit Singh Bedi, Csaba Szepesvári, and Mengdi Wang. Variational policy gradient method for reinforcement learning with general utilities. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, pages 4572–4583, 2020a.

Junyu Zhang, Chengzhuo Ni, Csaba Szepesvari, Mengdi Wang, et al. On the convergence and sample efficiency of variance-reduced policy gradient method. In *Advances in Neural Information Processing Systems*, volume 34, pages 2228–2240, 2021a.

Junzi Zhang, Jongho Kim, Brendan O'Donoghue, and Stephen Boyd. Sample efficient reinforcement learning with REINFORCE. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 10887–10895, 2021b.

Kaiqing Zhang, Bin Hu, and Tamer Basar. Policy optimization for $\mathcal{H}_2$ linear control with $\mathcal{H}_\infty$ robustness guarantee: Implicit regularization and global convergence. In *Learning for Dynamics and Control*, pages 179–190. PMLR, 2020b.

Kaiqing Zhang, Sham Kakade, Tamer Başar, and Lin Yang. Model-based multi-agent RL in zero-sum Markov games with near-optimal sample complexity. *Advances in Neural Information Processing Systems*, 33, 2020c.

Kaiqing Zhang, Alec Koppel, Hao Zhu, and Tamer Baş. Global convergence of policy gradient methods to (almost) locally optimal policies. *SIAM Journal on Control and Optimization*, 58(6): 3586–3612, 2020d.

Runyu Zhang, Qinghua Liu, Huan Wang, Caiming Xiong, Na Li, and Yu Bai. Policy optimization for Markov games: Unified framework and faster convergence. *Advances in Neural Information Processing Systems*, 35, 2022a.

Runyu Zhang, Jincheng Mei, Bo Dai, Dale Schuurmans, and Na Li. On the global convergence rates of decentralized softmax gradient play in Markov potential games. *Advances in Neural Information Processing Systems*, 35:1923–1935, 2022b.

Runyu Zhang, Zhaolin Ren, and Na Li. Gradient play in stochastic games: stationary points, convergence, and sample complexity. *IEEE Transactions on Automatic Control*, 2024a.

Shenao Zhang, Donghan Yu, Hiteshi Sharma, Ziyi Yang, Shuohang Wang, Hany Hassan, and Zhaoran Wang. Self-exploring language models: Active preference elicitation for online alignment. *arXiv preprint arXiv:2405.19332*, 2024b.

Tong Zhang. *Mathematical analysis of machine learning algorithms*. Cambridge University Press, 2023.

Xin Zhang, Jia Liu, and Zhengyuan Zhu. Taming convergence for asynchronous stochastic gradient descent with unbounded delay in non-convex learning. In *2020 59th IEEE Conference on Decision and Control (CDC)*, pages 3580–3585. IEEE, 2020e.

Yao Zhao, Rishabh Joshi, Tianqi Liu, Misha Khalman, Mohammad Saleh, and Peter J. Liu. SLiC-HF: Sequence likelihood calibration with human feedback. *arXiv preprint arXiv:2305.10425*, 2023.

Yulai Zhao, Yuandong Tian, Jason Lee, and Simon Du. Provably efficient policy optimization for two-player zero-sum Markov games. In *International Conference on Artificial Intelligence and Statistics*, pages 2736–2761. PMLR, 2022.

Banghua Zhu, Michael Jordan, and Jiantao Jiao. Principled reinforcement learning with human feedback from pairwise or k-wise comparisons. In *International Conference on Machine Learning*, pages 43037–43067. PMLR, 2023.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.