Journal of Machine Learning Research 25 (2024) 1-48 Submitted 10/21; Revised 12/22 & 10/23; Published 1/24

Fast Policy Extragradient Methods for Competitive Games with Entropy Regularization

Shicong Cen SHICONGC@ANDREW.CMU.EDU
Department of Electrical and Computer Engineering
Carnegie Mellon University

Yuting Wei YTWEI@WHARTON.UPENN.EDU
Department of Statistics and Data Science, The Wharton School
University of Pennsylvania

Yuejie Chi YUEJIECHI@CMU.EDU
Department of Electrical and Computer Engineering
Carnegie Mellon University

Editor: Csaba Szepesvari

Abstract

This paper investigates the problem of computing the equilibrium of competitive games in the form of two-player zero-sum games, which is often modeled as a constrained saddle-point optimization problem with probability simplex constraints. Despite recent efforts in understanding the last-iterate convergence of extragradient methods in the unconstrained setting, the theoretical underpinnings of these methods in the constrained settings, especially those using multiplicative updates, remain highly inadequate, even when the objective function is bilinear. Motivated by the algorithmic role of entropy regularization in single-agent reinforcement learning and game theory, we develop provably efficient extragradient methods to find the quantal response equilibrium (QRE)—which are solutions to zero-sum two-player matrix games with entropy regularization—at a linear rate. The proposed algorithms can be implemented in a decentralized manner, where each player executes symmetric and multiplicative updates iteratively using its own payoff without observing the opponent's actions directly. In addition, by controlling the knob of entropy regularization, the proposed algorithms can locate an approximate Nash equilibrium of the unregularized matrix game at a sublinear rate without assuming the Nash equilibrium to be unique. Our methods also lead to efficient policy extragradient algorithms for solving (entropy-regularized) zero-sum Markov games at similar rates. All of our convergence rates are nearly dimension-free, which are independent of the size of the state and action spaces up to logarithm factors, highlighting the positive role of entropy regularization for accelerating convergence.

Keywords: zero-sum Markov game, matrix game, entropy regularization, global convergence, multiplicative updates, no-regret learning, extragradient methods

1. Introduction

Finding the equilibrium of competitive games, which can be viewed as constrained saddle-point optimization problems with probability simplex constraints, lies at the heart of modern machine learning and decision making paradigms such as Generative Adversarial Networks

(GANs) (Goodfellow et al., 2020), competitive reinforcement learning (RL) (Littman, 1994), game theory (Shapley, 1953), adversarial training (Mertikopoulos et al., 2018b), to name a few.

In this paper, we study one of the most basic forms of competitive games, namely two-player zero-sum games, in both the matrix setting and the Markov setting. 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.

1.1 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 paper, 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.

1.2 Our 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) 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 2.1 for matrix games and Section 3.1 for Markov games, respectively.

- *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, 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

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

  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 $\widetilde{\mathcal{O}}\left(\|A\|_\infty\tau^{-1}\log(1/\epsilon)\right)$.

- *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

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

  iterations with optimized learning rates, again independent of the size of the action spaces up to logarithmic factors. Unlike prior literature (Daskalakis and Panageas,

Equilibrium type	Method	Convergence rate	Dimension-free	Require unique NE
ϵ-QRE	PU & OMWU **(this work)**	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 work)**	sublinear	yes	no

Table 1: 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.

2019; Wei et al., 2021a), our last-iterate convergence guarantee does not require the NE to be unique.

- *No-regret learning of entropy-regularized OMWU.* We further establish that under a decaying learning rate, the proposed OMWU method achieves a *logarithmic* regret for the entropy-regularized matrix game — on the order of $\mathcal{O}(\log T)$ — even when only one player follows the algorithm against arbitrary plays of the opponent. By setting τ appropriately, this translates to a regret of $\mathcal{O}((T\log T)^{1/2})$ for the unregularized matrix game, therefore matching the regret in Rakhlin and Sridharan (2013) without the need of mixing in an auxiliary uniform distribution for exploration.

- *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 $\widetilde{\mathcal{O}}\left(\frac{1}{\tau(1-\gamma)^2}\log^2\left(\frac{1}{\epsilon}\right)\right)$ 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 $\widetilde{\mathcal{O}}\left(\frac{1}{(1-\gamma)^3\epsilon}\right)$ iterations, which is independent of the dimension of the state-action space up to logarithmic factors.

To the best of our knowledge, our paper 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 and Table 2 provide detailed comparisons of the proposed methods with prior arts for solving

Equilibrium type	Method	Convergence rate	Dimension-free rate	Symmetric updates	Last-iterate convergence
ϵ-QRE	**(this work)**	linear	yes	yes	yes
ϵ-NE	(Perolat et al., 2015)	sublinear	no	no	yes
	(Zhao et al., 2022)	sublinear	no	no	yes
	(Daskalakis et al., 2020)	sublinear	no	no	no
	(Wei et al., 2021b)	sublinear	no	yes	yes
	(this work)	sublinear	yes	yes	yes

Table 2: Comparisons of the proposed policy extragradient method with competitive algorithms for finding ϵ-NE of two-player zero-sum Markov games. We note that the convergence rates in the prior arts all depend on various notions of concentrability coefficient and therefore not dimension-free. Desirable features in the last three columns are highlighted in blue.

competitive games. Our results highlight the positive role of entropy regularization for accelerating convergence and safeguarding against imperfect payoff information in competitive games.

1.3 Related works

Our work lies at the intersection of saddle-point optimization, game theory, and reinforcement learning. In what follows, we discuss a few topics that are closely related to ours.

Unregularized matrix game. Freund and Schapire (1999) showed that many standard methods such as GDA and MWU have a converging average duality gap at the rate of $O(1/\sqrt{T})$, which is improved to $O(1/T)$ 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. In terms of no-regret analysis, Rakhlin and Sridharan (2013) established a no-regret learning rate of

$O(\log T/T^{1/2})$ with an auxiliary mixing of a uniform distribution at each update, which is later improved to $O(1/T^{1/2})$ in Kangarshahi et al. (2018) with a slightly different algorithm.

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. (2020) 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 RL and games. In single-agent RL, the role of entropy regularization as an algorithmic mechanism to encourage exploration and accelerate convergence has been investigated extensively (Neu et al., 2017; Geist et al., 2019; Mei et al., 2020; Cen et al., 2022b; Lan, 2022; Zhan et al., 2021). Turning to the game setting, entropy regularization is 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); Cen et al. (2022a).

Zero-sum Markov games. There have been a significant recent interest in developing provably efficient self-play algorithms for Markov games, including model-based algorithms (Perolat et al., 2015; Sidford et al., 2020; Zhang et al., 2020; Li et al., 2022; Cui and Yang, 2021), value-based algorithms (Bai and Jin, 2020; Xie et al., 2020; Mao et al., 2022), and policy-based algorithms which either assumes access to gradient (Daskalakis et al., 2020), Bellman operator (Wei et al., 2021b), or both (Zhao et al., 2022). Our approach can be regarded as a policy-based algorithm to approximate value iteration, which can be implemented in a decentralized manner with symmetric and multiplicative updates from both players, and the iteration complexity is almost independent of the size of the state-action space. The iteration complexities in prior works (Perolat et al., 2015; Daskalakis et al., 2020; Wei et al., 2021b; Zhao et al., 2022) depend on various notions of concentrability coefficient and therefore can scale quite pessimistically with the problem dimension. In addition, while the last-iterate convergence guarantees in (Perolat et al., 2015; Daskalakis et al., 2020; Zhao et al., 2022) are applicable to the duality gap, Wei et al. (2021b) proves the last-iterate convergence in terms of the Euclidean distance to NE, together with an average convergence in terms of the duality gap.

It is worth mentioning that the study of model-based and value-based algorithms typically focuses on the statistical issues in terms of sample complexity under the generative

model or the online model of data collection; on the other end, the study of policy-based algorithms highlights the optimization issues by sidestepping the statistical issues using exact gradient evaluations, and later translating to sample complexity guarantees by leveraging model-based or value-based policy evaluation algorithms. Indeed, after the appearance of the initial version of this paper, Chen et al. (2021) has built on our algorithm to develop a sample-efficient version of policy extragradient methods in the online setting using bandit feedback.

1.4 Notation

We denote by $\Delta(\mathcal{A})$ the probability simplex over the set $\mathcal{A}$. We overload the functions such as $\log(\cdot)$ and $\exp(\cdot)$ to take vector inputs with the understanding that the function is applied in an entrywise manner. For instance, given any vector $z = [z_i]_{1\le i\le n} \in \mathbb{R}^n$, the notation $\exp(z)$ denotes $\exp(z) \coloneqq [\exp(z_i)]_{1\le i\le n}$; other functions are defined analogously. Given two probability distributions μ and μ' over $\mathcal{A}$, the KL divergence from μ' to μ is defined by $\mathsf{KL}\big(\mu \,\|\, \mu'\big) \coloneqq \sum_{a\in\mathcal{A}} \mu(a) \log \frac{\mu(a)}{\mu'(a)}$. 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 $\mathbf{1}$.

2. Zero-sum matrix games with entropy regularization

In this section, 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 paper, let $\mathcal{A} = \{1, \ldots, m\}$ and $\mathcal{B} = \{1, \ldots, n\}$ be the action spaces of each player. The proofs for this section are collected in Appendix A.

2.1 Background and problem formulation

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

$$\max_{\mu\in\Delta(\mathcal{A})} \min_{\nu\in\Delta(\mathcal{B})} f(\mu,\nu) \coloneqq \mu^\top A \nu, \tag{1}$$

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 von Neumann (1928) that the max and min operators in (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 (1) if

$$f(\mu^\star, \nu) \ge f(\mu^\star, \nu^\star) \ge f(\mu, \nu^\star) \qquad \text{for all } (\mu,\nu) \in \Delta(\mathcal{A}) \times \Delta(\mathcal{B}). \tag{2}$$

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):

$$\max_{\mu\in\Delta(\mathcal{A})}\min_{\nu\in\Delta(\mathcal{B})} f_\tau(\mu,\nu) \coloneqq \mu^\top A\nu + \tau\mathcal{H}(\mu) - \tau\mathcal{H}(\nu), \tag{3}$$

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 (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:

$$\begin{cases} \mu_\tau^\star(a) = \frac{\exp([A\nu_\tau^\star]_a/\tau)}{\sum_{a=1}^m \exp([A\nu_\tau^\star]_a/\tau)} \propto \exp([A\nu_\tau^\star]_a/\tau), & \text{for all } a\in\mathcal{A}, \\ \nu_\tau^\star(b) = \frac{\exp(-[A^\top\mu_\tau^\star]_b/\tau)}{\sum_{b=1}^n \exp(-[A^\top\mu_\tau^\star]_b/\tau)} \propto \exp(-[A^\top\mu_\tau^\star]_b/\tau), & \text{for all } b\in\mathcal{B}. \end{cases} \tag{4}$$

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 (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 (3), which is specified in (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

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

The duality gap of the entropy-regularized matrix game (3) at $\zeta = (\mu,\nu)$ is defined as

$$\mathsf{DualGap}_\tau(\zeta) = \max_{\mu'\in\Delta(\mathcal{A})} f_\tau(\mu',\nu) - \min_{\nu'\in\Delta(\mathcal{B})} f_\tau(\mu,\nu') \tag{5}$$

which is clearly nonnegative and $\mathsf{DualGap}_\tau(\zeta_\tau^\star) = 0$. Similarly, let the optimality gap of the entropy-regularized matrix game (3) at $\zeta = (\mu,\nu)$ be $\mathsf{OptGap}(\zeta) = \big|f_\tau(\mu,\nu) - f_\tau(\mu_\tau^\star,\nu_\tau^\star)\big|$.

2.2 Proposed extragradient methods: PU and OMWU

To begin, assume we are given a pair of policies $z_1\in\Delta(\mathcal{A})$, $z_2\in\Delta(\mathcal{B})$ 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 (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

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

where η is the learning rate, or equivalently

$$\begin{cases} \mu^{(t+1)}(a) \propto \mu^{(t)}(a)^{1-\eta\tau} \exp(\eta[Az_2]_a), & \text{for all } a \in \mathcal{A}, \\ \nu^{(t+1)}(b) \propto \nu^{(t)}(b)^{1-\eta\tau} \exp(-\eta[A^\top z_1]_b), & \text{for all } b \in \mathcal{B}. \end{cases} \tag{6}$$

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:

$$\text{Implicit Update:} \quad \begin{cases} \mu^{(t+1)}(a) \propto \mu^{(t)}(a)^{1-\eta\tau} \exp(\eta[A\nu^{(t+1)}]_a), & \text{for all } a \in \mathcal{A}, \\ \nu^{(t+1)}(b) \propto \nu^{(t)}(b)^{1-\eta\tau} \exp(-\eta[A^\top \mu^{(t+1)}]_b), & \text{for all } b \in \mathcal{B}. \end{cases} \tag{7}$$

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)). 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 (7) satisfy*

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

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 (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 1 and Algorithm 2, 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. (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).

Algorithm 1: 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

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

- The update of $(\mu^{(t+1)}, \nu^{(t+1)})$ then mimics the implicit update (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 2 reduces to the method analyzed in Rakhlin and Sridharan (2013); Daskalakis and Panageas (2019); Wei et al. (2021a) without entropy regularization.

2.3 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 (3). Its proof can be found in Section A.2.

Theorem 2 (Last-iterate convergence of PU and OMWU) *Suppose that the learning rates $\eta_t = \eta = \eta_{\mathsf{PU}}$ of PU in Algorithm 1 and $\eta_t = \eta = \eta_{\mathsf{OMWU}}$ of OMWU in Algorithm 2 satisfy*

$$0 < \eta_{\mathsf{PU}} \le \frac{1}{\tau + 2\|A\|_\infty}, \quad \text{and} \quad 0 < \eta_{\mathsf{OMWU}} \le \min\left\{\frac{1}{2\tau + 2\|A\|_\infty}, \frac{1}{4\|A\|_\infty}\right\}. \tag{8}$$

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:**

$$\max\left\{\mathsf{KL}\big(\zeta_\tau^\star \,\|\, \zeta^{(t)}\big),\ \tfrac{1}{2}\mathsf{KL}\big(\zeta_\tau^\star \,\|\, \bar{\zeta}^{(t+1)}\big)\right\} \le (1-\eta\tau)^t \mathsf{KL}\big(\zeta_\tau^\star \,\|\, \zeta^{(0)}\big), \tag{9a}$$

Algorithm 2: 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

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

$$\left\| \log \frac{\zeta^{(t)}}{\zeta_\tau^\star} \right\|_\infty \le 2(1-\eta\tau)^t \left\| \log \frac{\zeta^{(0)}}{\zeta_\tau^\star} \right\|_\infty + \frac{8\|A\|_\infty}{\tau} (1-\eta\tau)^{t/2} \mathsf{KL}\big(\zeta_\tau^\star \,\|\, \zeta^{(0)}\big)^{1/2}. \tag{9b}$$

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

$$\mathsf{OptGap}_\tau(\bar{\zeta}^{(t)}) \le 3\eta^{-1}(1-\eta\tau)^t \mathsf{KL}\big(\zeta_\tau^\star \,\|\, \zeta^{(0)}\big), \tag{9c}$$

$$\mathsf{DualGap}_\tau(\bar{\zeta}^{(t)}) \le \left(\eta^{-1} + 2\tau^{-1}\|A\|_\infty^2\right)(1-\eta\tau)^{t-1} \mathsf{KL}\big(\zeta_\tau^\star \,\|\, \zeta^{(0)}\big). \tag{9d}$$

Remark 3 *To further understand the term* $\mathsf{KL}\big(\zeta_\tau^\star \,\|\, \zeta^{(0)}\big)$ *in* (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 4 *Similar results continue to hold even when the two players use different regularization parameters* $\tau_\mu, \tau_\nu > 0$ *in* (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 2 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. (8)), and hence leads to an equivalent level of computational efficiency. Compared with the vast literature analyzing the average-iterate performance of variants of extragradient methods, 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 (9a) tells that it is sufficient to take

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

  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 $(1+\|A\|_\infty/\tau)\log(1/\epsilon)$ (modulo log factors), which only depends on the ratio $\|A\|_\infty/\tau$.[1]

- **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. (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. (9c)) and the duality gap (cf. (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 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.

Remark 5 *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. In particular, the mirror prox algorithm proceed by applying*

$$\begin{cases}\mu^{(t+1)}(a)\propto\mu^{(t)}(a)\bar{\mu}^{(t+1)}(a)^{-\eta\tau}\exp(\eta[A\bar{\nu}^{(t+1)}]_a)\\ \nu^{(t+1)}(b)\propto\nu^{(t)}(b)\bar{\nu}^{(t+1)}(b)^{-\eta\tau}\exp(-\eta[A^\top\bar{\mu}^{(t+1)}]_b)\end{cases}, \tag{10}$$

while PU restrict the extragradient step to the bilinear part and yields the update:

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

This seemingly small, but important, difference leads to a more concise closed-form update rule and a tractable analysis. It is of great interest to see if similar last-iterate convergence

1. We remark that evaluating the matrix-vector product $A\nu$ and $A^\top\mu$ still takes up to $\mathcal{O}(|\mathcal{A}|^2)$ computation cost per iteration.

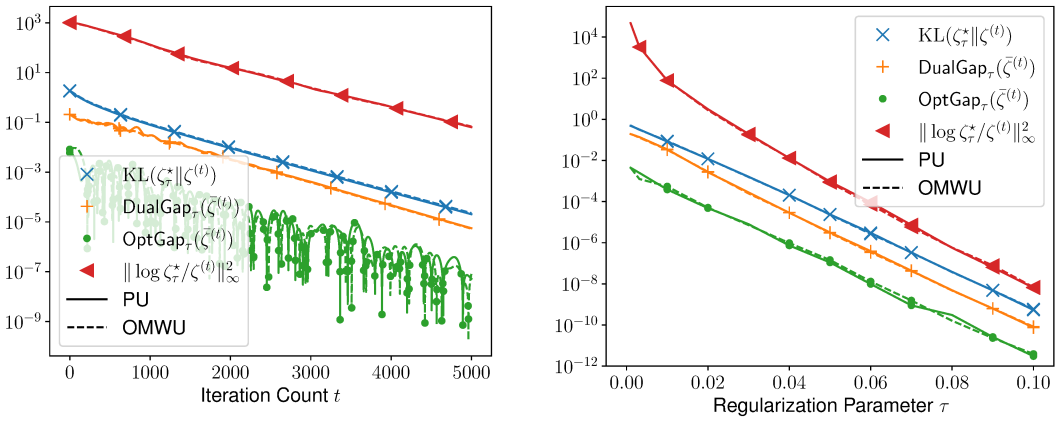

Figure 1: Performance illustration of the PU and OMWU methods for solving entropy-regularized matrix games with $|\mathcal{A}| = |\mathcal{B}| = 100$, where the entries of the payoff matrix A is generated independently from the uniform distribution on $[-1, 1]$. The learning rates are fixed as $\eta = 0.1$. 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.

guarantees can be established to the mirror prox update rule (10)*, which to the best of knowledge, is not available in the literature.*[2]

Last-iterate convergence to approximate NE. The entropy-regularized matrix game can be thought as a smooth surrogate of the unregularized matrix game (1); in particular, it is possible to find an ϵ-NE by setting τ sufficiently small in (3). According to (Zhang et al., 2020, Definition 2.1), a policy pair $\zeta = (\mu, \nu)$ is an ϵ-NE if it satisfies

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

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

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

in view of the boundedness of the Shannon entropy $\mathcal{H}(\cdot)$. Theorem 9 (cf. (9d)) also ensures that our proposed algorithms find an approximate QRE $\bar{\zeta}^{(T)}$ such that $\mathsf{DualGap}_\tau(\bar{\zeta}^{(T)}) \le \epsilon/2$

2. It was brought to our attention during the review process that an unpublished arXiv note Titov (2021), which was posted after our initial arXiv version, claimed a similar last-iterate convergence guarantee for some restarted variant of mirror prox. However, we note that Theorem 1 in the note exhibited a different form from its reference (Stonyakin et al., 2020, Theorem 4.8) and as such the proof of the main result of the note has a gap. Further scrutiny and efforts might be needed to establish the performance of vanilla mirror prox for solving the entropy-regularized matrix game.

after taking $T = \widetilde{O}\left(\frac{1}{\eta\epsilon}\right)$ iterations, which is no more than

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

iterations with optimized learning rates. It follows immediately that

$$\begin{aligned}&\mathsf{DualGap}(\bar{\zeta}^{(T)})\\ &\leq \mathsf{DualGap}_\tau(\bar{\zeta}^{(T)}) + \max_{\mu',\nu'}\left|f_\tau(\mu',\bar{\nu}^{(T)}) - f_\tau(\bar{\mu}^{(T)},\nu') - (f(\mu',\bar{\nu}^{(T)}) - f(\bar{\mu}^{(T)},\nu'))\right| \leq \epsilon, \quad (11)\end{aligned}$$

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 (1) to be unique. See Table 1 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

$$\mu^{(t+1)}(a) \propto \mu^{(t)}(a)^{1-\eta\tau}\exp(\eta[A\nu]_a). \quad (12)$$

In this case, Theorem 2 can be established in exactly the same fashion by restricting attention only to the updates of $\mu^{(t)}$.

2.4 No-regret learning of entropy-regularized OMWU

Besides convergence to equilibria, in game-theoretical settings, it is often desirable to design and implement no-regret algorithms, which are capable of providing black-box guarantees over arbitrary sequences played by the opponent (Cesa-Bianchi and Lugosi, 2006; Rakhlin and Sridharan, 2013). Therefore, no-regret algorithms provide a sort of robustness especially when operating in contested environments, when the opponent is potentially adversarial. Fortunately, it turns out that entropy regularization not only accelerates the convergence, but also enables no-regret learning somewhat "for free": it encourages exploration by putting a positive mass on every action, therefore guards against adversaries. Moreover, since algorithms that call the sampling oracle more than once per iteration often incurs a linear regret (Golowich et al., 2020), we focus on OMWU (Algorithm 2) and establish it as a no-regret algorithm.

No-regret learning of OMWU for the entropy-regularized matrix game. We begin by formally defining the notion of regret. Suppose that player 2 plays according to Algorithm 2 to update $\nu^{(t)}$ and $\bar{\nu}^{(t)}$ based on the received payoff sequence $\{A^\top \bar{\mu}^{(t)}\}$, $t = 0, 1, \ldots$, whose construction potentially is deviated from the update rule of Algorithm 2, and even adversarial. Let

$$f_\tau^{(t)}(\nu) = \bar{\mu}^{(t)\top} A\nu + \tau\mathcal{H}(\bar{\mu}^{(t)}) - \tau\mathcal{H}(\nu),$$

which is the regularized game value upon fixing the policy of player 1 as $\mu = \bar{\mu}^{(t)}$. The regret experienced by player 2 is then defined as

$$\mathsf{Regret}_\lambda(T) = \sum_{t=0}^{T} f_\tau^{(t)}(\bar{\nu}^{(t)}) - \min_{\nu\in\Delta(\mathcal{B})} \sum_{t=0}^{T} f_\tau^{(t)}(\nu), \tag{13}$$

which measures the gap between the actual performance and the performance in hindsight. The following theorem shows that with appropriate choices of the learning rate, OMWU achieves a logarithmic regret bound $O(\log T)$.

Theorem 7 *Suppose only one player (say, player 2) follows the entropy-regularized OMWU method in Algorithm 2. Setting the learning rate as $\eta_t = \frac{1}{(t+1)\tau}$ and the initialization policy as the uniform policy, i.e. $\nu^{(0)}(b) = 1/|\mathcal{B}|, \forall b \in \mathcal{B}$, the regret against any sequence $\{A^\top \bar{\mu}^{(t)}\}_{t=0}^T$ of play is bounded by*

$$\mathsf{Regret}_\lambda(T) \le \tau^{-1}(\log T + 1)(\tau \log|\mathcal{B}| + 5\left\|A\right\|_\infty)^2 + 4\left\|A\right\|_\infty.$$

Theorem 7 implies that the average regret satisfies

$$\frac{1}{T}\mathsf{Regret}_\lambda(T) \lesssim \frac{\log T}{T},$$

which goes to zero as T increases, therefore implies the entropy-regularized OMWU method is no-regret.

No-regret learning for the unregularized matrix game. Similar to earlier discussions, one can still hope to control the regret of the unregularized matrix game, by appropriately setting τ sufficiently small. It is easily seen that the regret of the unregularized matrix game is given by

$$\begin{aligned}
\mathsf{Regret}_0(T) &= \sum_{t=0}^{T} f_0^{(t)}(\bar{\nu}^{(t)}) - \min_{\nu\in\Delta(\mathcal{B})} \sum_{t=0}^{T} f_0^{(t)}(\nu) = \max_{\nu\in\Delta(\mathcal{B})} \sum_{t=0}^{T} \Big[f_0^{(t)}(\bar{\nu}^{(t)}) - f_0^{(t)}(\nu)\Big] \\
&\le \mathsf{Regret}_\lambda(T) + \max_{\nu\in\Delta(\mathcal{B})} \Big\{ -\tau \sum_{t=0}^{T} \mathcal{H}(\bar{\nu}^{(t)}) + \tau T\mathcal{H}(\nu)\Big\} \\
&\le \mathsf{Regret}_\lambda(T) + 2\tau T \log|\mathcal{B}| \\
&\le \tau^{-1}(\log T + 1)(\tau \log|\mathcal{B}| + 5\left\|A\right\|_\infty)^2 + 4\left\|A\right\|_\infty + 2\tau T \log|\mathcal{B}|.
\end{aligned}$$

Therefore, by setting $\tau = O((\log T/T)^{1/2})$, we can ensure that the regret with regard to the unregularized problem is bounded by $\mathsf{Regret}_0(T) = O((T\log T)^{1/2})$, which is on par

with the regret established in Rakhlin and Sridharan (2013). It is worthwhile to highlight that the OMWU method in (Rakhlin and Sridharan, 2013) requires blending in a uniform distribution every iteration to guarantee no-regret learning, while a similar blending is enabled in ours without extra algorithmic steps.

3. Zero-sum Markov games with entropy regularization

Leveraging the success of PU and OMWU in solving the entropy-regularized matrix games, this section extends our current analysis to solve the zero-sum two-player Markov game, which is again formulated as finding the equilibrium of a saddle-point optimization problem. We start by introducing its basic setup, along with the entropy-regularized Markov game, which will be followed by the proposed policy extragradient method with its theoretical guarantees. The proofs for this section are collected in Appendix B.

3.1 Background and problem formulation

Zero-sum two-player Markov game. We consider a discounted Markov Game (MG) which is defined as $\mathcal{M} = \{\mathcal{S}, \mathcal{A}, \mathcal{B}, P, r, \gamma\}$, with discrete state space $\mathcal{S}$, action spaces of two players $\mathcal{A}$ and $\mathcal{B}$, transition probability P, reward function $r : \mathcal{S} \times \mathcal{A} \times \mathcal{B} \to [0, 1]$ and discount factor $\gamma \in [0, 1)$. A policy $\mu : \mathcal{S} \to \Delta(\mathcal{A})$ (resp. $\nu : \mathcal{S} \to \Delta(\mathcal{B})$) defines how player 1 (resp. player 2) reacts to a given state s, where the probability of taking action $a \in \mathcal{A}$ (resp. $b \in \mathcal{B}$) is $\mu(a|s)$ (resp. $\nu(b|s)$). The transition probability kernel $P : \mathcal{S} \times \mathcal{A} \times \mathcal{B} \to \Delta(\mathcal{S})$ defines the dynamics of the Markov game, where $P(s'|s, a, b)$ specifies the probability of transiting to state s' from state s when the players take actions a and b respectively. The state value of a given policy pair (μ, ν) is evaluated by the expected discounted cumulative reward:

$$V^{\mu,\nu}(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t, b_t) \,\middle|\, s_0 = s\right],$$

where the trajectory $(s_0, a_0, b_0, s_1, \cdots)$ is generated by the MG $\mathcal{M}$ under the policy pair (μ, ν), starting from the state s_0. Similarly, the Q-function captures the expected discounted cumulative reward with an initial state s and initial action pair (a, b) for a given policy pair (μ, ν):

$$Q^{\mu,\nu}(s, a, b) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t, b_t) \,\middle|\, s_0 = s, a_0 = a, b_0 = b\right].$$

In a zero-sum game, one player seeks to maximize the value function while the other player wants to minimize it. The minimax game value on state s is defined by

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

Similarly, the minimax Q-function $Q^\star(s, a, b)$ is defined by

$$Q^\star(s, a, b) = r(s, a, b) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a,b)} V^\star(s'). \tag{14}$$

It is proved by Shapley (1953) that a pair of stationary policy $(\mu^\star, \nu^\star)$ attaining the minimax value on state s attains the minimax value on all states as well (Filar and Vrieze, 2012), and is called the NE of the MG.

Entropy-regularized zero-sum two-player Markov game. Motivated by entropy regularization in Markov decision processes (MDP) (Geist et al., 2019; Cen et al., 2022b), we consider an entropy-regularized variant of MG, where the value function is modified as

$$V_\tau^{\mu,\nu}(s) := \mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t\left(r(s_t,a_t,b_t) - \tau\log\mu(a_t|s_t) + \tau\log\nu(b_t|s_t)\right)\,\middle|\, s_0 = s\right], \tag{15}$$

where the quantity $\tau \geq 0$ denotes the regularization parameter, and the expectation is evaluated over the randomness of the transition kernel as well as the policies. The regularized Q-function $Q_\tau^{\mu,\nu}$ of a policy pair (μ,ν) is related to $V_\tau^{\mu,\nu}$ as

$$Q_\tau^{\mu,\nu}(s,a,b) = r(s,a,b) + \gamma\mathbb{E}_{s'\sim P(\cdot|s,a,b)}\left[V_\tau^{\mu,\nu}(s')\right]. \tag{16}$$

We will call $V_\tau^{\mu,\nu}$ and $Q_\tau^{\mu,\nu}$ the *soft value function* and *soft Q-function*, respectively. A policy pair $(\mu_\tau^\star,\nu_\tau^\star)$ is said to be the quantal response equilibrium (QRE) of the entropy-regularized MG, if its value attains the minimax value of the entropy-regularized MG over all states $s\in\mathcal{S}$, i.e.

$$V_\tau^\star(s) = \max_\mu\min_\nu V_\tau^{\mu,\nu}(s) = \min_\nu\max_\mu V_\tau^{\mu,\nu}(s) := V_\tau^{\mu_\tau^\star,\nu_\tau^\star}(s),$$

where $V_\tau^\star$ is called the optimal minimax soft value function, and similarly $Q_\tau^\star := Q_\tau^{\mu_\tau^\star,\nu_\tau^\star}$ is called the optimal minimax soft Q-function.

Goal. Our goal is to find the QRE of the entropy-regularized MG in a decentralized manner where the players only observe its own reward without accessing the opponent's actions, and leverage the QRE to find an approximate NE of the unregularized MG.

3.2 From value iteration to policy extragradient methods

Entropy-regularized value iteration. It is known that classical dynamic programming approaches such as value iteration can be extended to solve MG (Perolat et al., 2015), where each iteration amounts to solving a series of matrix games for each state. Similar to the single-agent case (Cen et al., 2022b), we can extend these approaches to solve the entropy-regularized MG. Setting the stage, let us introduce the per-state Q-value matrix $Q(s) := Q(s,\cdot,\cdot)\in\mathbb{R}^{|\mathcal{A}|\times|\mathcal{B}|}$ for every $s\in\mathcal{S}$, where the element indexed by the action pair (a,b) is $Q(s,a,b)$. Similarly, we define the per-state policies $\mu(s) := \mu(\cdot|s)\in\Delta(\mathcal{A})$ and $\nu(s) := \nu(\cdot|s)\in\Delta(\mathcal{B})$ for both players.

In parallel to the original Bellman operator, we denote the *soft Bellman operator* $\mathcal{T}_\tau$ as

$$\mathcal{T}_\tau(Q)(s,a,b) := r(s,a,b) + \gamma\mathbb{E}_{s'\sim P(\cdot|s,a,b)}\left[\max_{\mu(s')\in\Delta(\mathcal{A})}\min_{\nu(s')\in\Delta(\mathcal{B})} f_\tau\big(Q(s');\mu(s'),\nu(s')\big)\right], \tag{17}$$

where for each per-state Q-value matrix $Q(s)$, we introduce an entropy-regularized matrix game in the form of

$$\max_{\mu\in\Delta(\mathcal{A})}\min_{\nu\in\Delta(\mathcal{B})} f_\tau\big(Q(s);\mu(s),\nu(s)\big) := \mu(s)^\top Q(s)\nu(s) - \tau\mathcal{H}(\mu(s)) + \tau\mathcal{H}(\nu(s)).$$

Algorithm 3: Policy Extragradient Method applied to Value Iteration for Entropy-regularized Markov Game

initialization: $Q^{(0)} = 0$.

for $t = 0, 1, 2, \cdots, T_{\text{main}}$ **do**

Let $Q^{(t)}$ denote

$$Q^{(t)}(s,a,b) = r(s,a,b) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a,b)} V^{(t)}(s'). \tag{20}$$

Invoke PU (Algorithm 1) or OMWU (Algorithm 2) for T_{sub} iterations to solve the following entropy-regularized matrix game for every state s, where the initialization is set as uniform distributions:

$$\max_{\mu(s)\in\Delta(\mathcal{A})} \min_{\nu(s)\in\Delta(\mathcal{B})} f_\tau\big(Q^{(t)}(s); \mu(s), \nu(s)\big).$$

Return the last iterate $\bar{\mu}^{(t,T_{\text{sub}})}(s), \bar{\nu}^{(t,T_{\text{sub}})}(s)$.

Set $V^{(t+1)}(s) = f_\tau\big(Q^{(t)}(s); \bar{\mu}^{(t,T_{\text{sub}})}(s), \bar{\nu}^{(t,T_{\text{sub}})}(s)\big)$.

The entropy-regularized value iteration then proceeds as

$$Q^{(t+1)} = \mathcal{T}_\tau(Q^{(t)}), \tag{18}$$

where $Q^{(0)}$ is an initialization. By definition, the optimal minimax soft Q-function obeys $\mathcal{T}_\tau(Q_\tau^\star) = Q_\tau^\star$ and therefore corresponds to the fix point of the soft Bellman operator. Given the above entropy-regularized value iteration, the following lemma states its iterates contract linearly to the optimal minimax soft Q-function at a rate of the discount factor γ.

Proposition 2 *The entropy-regularized value iteration* (18) *converges at a linear rate, i.e.*

$$\|Q^{(t)} - Q_\tau^\star\|_\infty \le \gamma^t \|Q^{(0)} - Q_\tau^\star\|_\infty. \tag{19}$$

Approximate value iteration via policy extragradient methods. Proposition 2 suggests that the optimal minimax soft Q-function of the entropy-regularized MG can be found by solving a series of entropy-regularized matrix games induced by $\{Q^{(t)}\}_{t\ge 0}$ in (18), a task that can be accomplished by adopting the fast extragradient methods developed earlier. To proceed, we first define the following first-order oracle, which makes it rigorous that the proposed algorithm does not require access to the Q-function of the entire MG, but only its own single-agent Q-function when playing against the opponent's policy.

Definition 8 (first-order oracle for Markov games) *Given any policy pair $\mu(s), \nu(s)$ and Q-value matrix $Q(s)$ for any $s \in \mathcal{S}$, the first-order oracle returns*

$$[Q(s)\nu(s)]_a = \mathbb{E}_{b\sim\nu(s)}\left[Q(s,a,b)\right], \qquad \text{and} \qquad [Q(s)^\top \mu(s)]_b = \mathbb{E}_{a\sim\mu(s)}\left[Q(s,a,b)\right]$$

for any $a \in \mathcal{A}$ and $b \in \mathcal{B}$.

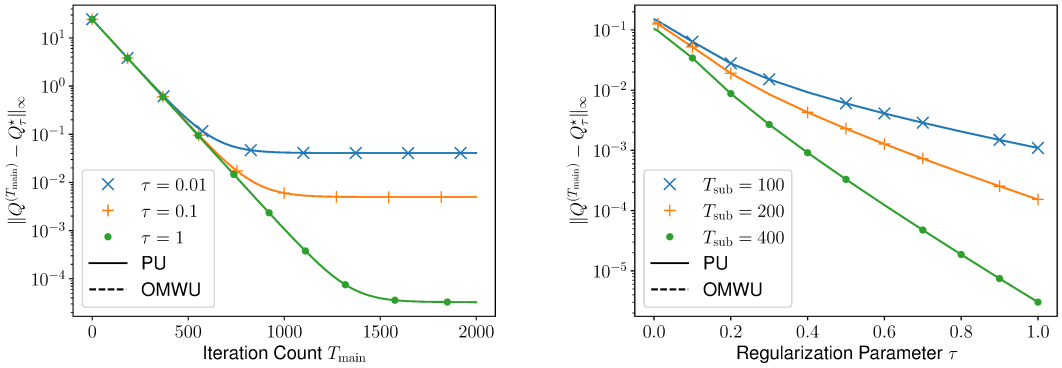

Figure 2: Performance illustration of Algorithm 3 for solving a random generated entropy-regularized Markov game with $|\mathcal{A}| = |\mathcal{B}| = 20$, $|\mathcal{S}| = 100$ and $\gamma = 0.99$. The learning rates of both players are fixed as $\eta = 0.005$. The left panel plots $\left\|Q^{(T_{\text{main}})} - Q_\tau^\star\right\|_\infty$ w.r.t. the iteration count T_{main} when $T_{\text{sub}} = 400$ under various entropy regularization parameters, while the right panel plots $\left\|Q^{(T_{\text{main}})} - Q_\tau^\star\right\|_\infty$ w.r.t. the regularization parameter τ when $T_{\text{main}} = 2000$ with different choices of T_{sub}. Due to their similar nature, PU and OMWU yield almost identical convergence behaviors and overlapping plots.

Algorithm 3 describes the proposed policy extragradient method. Encouragingly, by judiciously setting the number of iterations in both the outer loop (for updating the Q-value matrices) and the inner loop (for updating the QRE of the corresponding Q-value matrix), we are guaranteed to find the QRE of the entropy-regularized MG in a small number of iterations, as dictated by the following theorem.

Theorem 9 *Assume $|\mathcal{A}| \geq |\mathcal{B}|$ and $\tau \leq 1$. Setting $\eta_t = \eta = \frac{1-\gamma}{2(1+\tau(\log|\mathcal{A}|+1-\gamma))}$, the total iterations (namely, the product $T_{\text{main}} \cdot T_{\text{sub}}$) required for Algorithm 3 to achieve*

$$\left\|Q^{(T_{\text{main}})} - Q_\tau^\star\right\|_\infty \leq \epsilon$$

is at most

$$O\left(\frac{(\log|\mathcal{A}| + 1/\tau)}{(1-\gamma)^2}\left(\log\frac{\log|\mathcal{A}|}{(1-\gamma)\epsilon}\right)^2\right).$$

Theorem 9 ensures that within $\widetilde{O}\left(\frac{1}{\tau(1-\gamma)^2}\log^2\left(\frac{1}{\epsilon}\right)\right)$ iterations, Algorithm 3 finds a close approximation of the optimal minimax soft Q-function $Q_\tau^\star$ in an entrywise manner to a prescribed accuracy ϵ. Remarkably, the iteration complexity is independent of the dimensions of the state space and the action space (up to log factors). In addition, the iteration complexity becomes smaller when the amount of regularization increases.

Figure 2 illustrates the performance of Algorithm 3 for solving a randomly generated entropy-regularized Markov game with $|\mathcal{A}| = |\mathcal{B}| = 20$, $|\mathcal{S}| = 100$ and $\gamma = 0.99$ with

varying choices of T_{main}, T_{sub} and τ. Here, the transition probability kernel and the reward function are generated as follows. For each state-action pair (s,a,b), we randomly select 10 states to form a set $\mathcal{S}_{s,a,b}$, and set $P(s'|s,a,b) \propto U_{s,a,b,s'}$ if $s' \in \mathcal{S}_{s,a}$, and 0 otherwise, where $\{U_{s,a,b,s'} \sim U[0,1]\}$ are drawn independently from the uniform distribution over $[0,1]$. The reward function is generated by $r(s,a,b) \sim U_{s,a,b} \cdot U_s$, where $U_{s,a,b}$ and U_s are drawn independently from the uniform distribution over $[0,1]$. It is seen that the convergence speed of the ℓ_∞ error on $\left\|Q^{(T_{\mathrm{main}})} - Q_\tau^\star\right\|_\infty$ improves as we increase the regularization parameter τ, which corroborates our theory.

3.3 Last-iterate convergence to approximate NE

Similar to the case of matrix games, solving the entropy-regularized MG provides a viable strategy to find an ϵ-approximate NE of the unregularized MG, where the optimality of a policy pair is typically gauged by the duality gap (Zhang et al., 2020). To begin, define the duality gap of the entropy-regularized MG at a policy pair $\zeta = (\mu, \nu)$ as

$$\mathsf{DualGap}_\tau^{\mathsf{markov}}(\zeta) = \max_{\mu' \in \Delta(\mathcal{A})^{|\mathcal{S}|}} V_\tau^{\mu',\nu}(\rho) - \min_{\nu' \in \Delta(\mathcal{B})^{|\mathcal{S}|}} V_\tau^{\mu,\nu'}(\rho), \tag{21}$$

where ρ is an arbitrary distribution over the state space $\mathcal{S}$, and $V_\tau^{\mu,\nu}(\rho) := \mathbb{E}_{s\sim\rho}\left[V_\tau^{\mu,\nu}(s)\right]$.[3] Similarly, the duality gap of a policy pair $\zeta = (\mu,\nu)$ for the unregularized MG is defined as

$$\mathsf{DualGap}^{\mathsf{markov}}(\zeta) = \max_{\mu' \in \Delta(\mathcal{A})^{|\mathcal{S}|}} V^{\mu',\nu}(\rho) - \min_{\nu' \in \Delta(\mathcal{B})^{|\mathcal{S}|}} V^{\mu,\nu'}(\rho),$$

where $V^{\mu,\nu}(\rho) := \mathbb{E}_{s\sim\rho}\left[V^{\mu,\nu}(s)\right]$. Following Zhang et al. (2020), a pair of policy ζ is said to be ϵ-approximate NE if $\mathsf{DualGap}^{\mathsf{markov}}(\zeta) \le \epsilon$. Notice that for any policy pair (μ,ν), it is straightforward that

$$|V_\tau^{\mu,\nu}(\rho) - V^{\mu,\nu}(\rho)| \le \frac{\tau}{1-\gamma}(\log|\mathcal{A}| + \log|\mathcal{B}|).$$

Encouragingly, the following corollary ensures that Algorithm 3 yields a policy pair with ϵ-optimal duality gap for the entropy-regularized MG.

Corollary 10 *Assume $|\mathcal{A}| \ge |\mathcal{B}|$ and $\tau \le 1$. Setting $\eta_t = \eta = \frac{1-\gamma}{2(1+\tau(\log|\mathcal{A}|+1-\gamma))}$, Algorithm 3 takes no more than $\widetilde{O}\left(\frac{1}{\tau(1-\gamma)^2}\log^2\left(\frac{1}{\epsilon}\right)\right)$ iterations to achieve $\mathsf{DualGap}_\tau^{\mathsf{markov}}(\zeta) \le \epsilon$.*

With Corollary 10 in place, setting the regularization parameter sufficiently small, i.e. $\tau = O\big(\frac{(1-\gamma)\epsilon}{\log|\mathcal{A}|}\big)$, and invoking similar discussions as (11) allows us to find an ϵ-approximate NE of the unregularized MG within

$$\widetilde{O}\left(\frac{1}{(1-\gamma)^3\epsilon}\right)$$

iterations. See Table 2 for further comparisons with Perolat et al. (2015); Wei et al. (2021b); Daskalakis et al. (2020); Zhao et al. (2022). To the best of our knowledge, the proposed method is the only one that simultaneously possesses symmetric updates, problem-independent rates, and last-iterate convergence.

3. For notation simplicity, we omitted the dependency with ρ in $\mathsf{DualGap}_\tau^{\mathsf{markov}}(\zeta)$.

4. Conclusions

This paper develops provably efficient policy extragradient methods (PU and OMWU) for entropy-regularized matrix games and Markov 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 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)? A recent work by Cen et al. (2023) partially answered this question under exact gradient evaluation. Can we generalize the proposed algorithms to handle more general regularization terms, similar to what has been accomplished in the single-agent setting (Lan, 2022; Zhan et al., 2021)? Can we generalize the proposed algorithm to other type of games (Ao et al., 2023)? We leave the answers to future work.

Acknowledgments

S. Cen and Y. Chi are supported in part by the grants ONR N00014-18-1-2142 and N00014-19-1-2404, ARO W911NF-18-1-0303, NSF CCF-1901199, CCF-2007911 and CCF-2106778. S. Cen is also gratefully supported by Wei Shen and Xuehong Zhang Presidential Fellowship and Nicholas Minnici Dean's Graduate Fellowship in Electrical and Computer Engineering at Carnegie Mellon University, and JP Morgan Chase AI Research PhD Fellowship. Y. Wei is supported in part by the NSF grants CCF-2007911, DMS-2147546/2015447, CCF-2106778, CAREER award DMS-2143215, and the Google Research Scholar Award.

References

Ruicheng Ao, Shicong Cen, and Yuejie Chi. Asynchronous gradient play in zero-sum multi-agent games. In *The Eleventh International Conference on Learning Representations*, 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.

Yu Bai and Chi Jin. Provable self-play algorithms for competitive reinforcement learning. In *International Conference on Machine Learning*, pages 551–560. PMLR, 2020.

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.

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 61st Conference on Decision and Control (CDC)*, pages 2833–2838. 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 Shaolei Du, and Lin Xiao. Faster last-iterate convergence of policy optimization in zero-sum Markov games. In *The Eleventh International Conference on Learning Representations*, 2023.

Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.

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.

Qiwen Cui and Lin F Yang. Minimax sample complexity for turn-based stochastic game. In *Uncertainty in Artificial Intelligence*, pages 1496–1504. PMLR, 2021.

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 *Innovations in Theoretical Computer Science*, 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.

Jerzy Filar and Koos Vrieze. *Competitive Markov decision processes*. Springer Science & Business Media, 2012.

Yoav Freund and Robert E Schapire. Adaptive game playing using multiplicative weights. *Games and Economic Behavior*, 29(1-2):79–103, 1999.

Matthieu Geist, Bruno Scherrer, and Olivier Pietquin. A theory of regularized Markov decision processes. In *International Conference on Machine Learning*, pages 2160–2169, 2019.

Noah Golowich, Sarath Pattathil, and Constantinos Daskalakis. Tight last-iterate convergence rates for no-regret learning in multi-player games. *Advances in Neural Information Processing Systems*, 33, 2020.

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.

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.

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*, 33, 2020.

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.

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

Galina M Korpelevich. The extragradient method for finding saddle points and other problems. *Matecon*, 12:747–756, 1976.

Guanghui Lan. Policy mirror descent for reinforcement learning: Linear convergence, new sampling complexity, and generalized problem classes. *Mathematical programming*, pages 1–48, 2022.

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.

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.

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.

Michael L Littman. Markov games as a framework for multi-agent reinforcement learning. In *Machine Learning Proceedings*, pages 157–163. Elsevier, 1994.

Weichao Mao, Lin Yang, Kaiqing Zhang, and Tamer Basar. On improving model-free algorithms for decentralized multi-agent reinforcement learning. In *International Conference on Machine Learning*, pages 15007–15049. PMLR, 2022.

Richard D McKelvey and Thomas R Palfrey. Quantal response equilibria for normal form games. *Games and economic behavior*, 10(1):6–38, 1995.

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

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.

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.

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.

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

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.

Sasha Rakhlin and Karthik Sridharan. Optimization, learning, and games with predictable sequences. *Advances in Neural Information Processing Systems*, 26, 2013.

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.

Lloyd S Shapley. Stochastic games. *Proceedings of the National Academy of Sciences*, 39 (10):1095–1100, 1953.

Aaron Sidford, Mengdi Wang, Lin Yang, and Yinyu Ye. Solving discounted stochastic two-player games with near-optimal time and sample complexity. In *International Conference on Artificial Intelligence and Statistics*, pages 2992–3002. PMLR, 2020.

Fedor Stonyakin, Alexander Tyurin, Alexander Gasnikov, Pavel Dvurechensky, Artem Agafonov, Darina Dvinskikh, Mohammad Alkousa, Dmitry Pasechnyuk, Sergei Artamonov, and Victorya Piskunova. Inexact relative smoothness and strong convexity for optimization and variational inequalities by inexact model. *arXiv preprint arXiv:2001.09013*, 2020.

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.

Alexander A Titov. Algorithm for solving variational inequalities with relatively strongly monotone operators. *arXiv preprint arXiv:2106.01442*, 2021.

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.

J von Neumann. Zur theorie der gesellschaftsspiele. *Mathematische annalen*, 100(1):295–320, 1928.

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.

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.

Abhay Yadav, Sohil Shah, Zheng Xu, David Jacobs, and Tom Goldstein. Stabilizing adversarial nets with prediction methods. In *International Conference on Learning Representations*, 2018.

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.

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. *arXiv preprint arXiv:2105.11066*, 2021.

Kaiqing Zhang, Sham Kakade, Tamer Basar, 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, 2020.

Yulai Zhao, Yuandong Tian, Jason D Lee, and Simon S Du. Provably efficient policy gradient methods for two-player zero-sum Markov games. In *International Conference on Artificial Intelligence and Statistics*, 2022.

Appendix A. Analysis for entropy-regularized matrix games

Before embarking on the main proof, it is useful to first consider the update rule (6) that underlies both PU and OMWU, which is reproduced below for convenience:

$$\begin{cases} \mu^{(t+1)}(a) \propto \mu^{(t)}(a)^{1-\eta\tau} \exp(\eta[Az_2]_a), & \text{for all } a \in \mathcal{A}, \\ \nu^{(t+1)}(b) \propto \nu^{(t)}(b)^{1-\eta\tau} \exp(-\eta[A^\top z_1]_b), & \text{for all } b \in \mathcal{B}, \end{cases} \tag{22}$$

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.1.

Lemma 11 *Denote $\zeta^{(t)} = (\mu^{(t)}, \nu^{(t)})$ and $\zeta(z) = (z_1, z_2)$. The update rule (22) satisfies:*

$$\left\langle \log \mu^{(t+1)} - (1-\eta\tau)\log \mu^{(t)} - \eta\tau \log \mu_\tau^\star, z_1 - \mu_\tau^\star \right\rangle = \eta(\mu_\tau^\star - z_1)^\top A(\nu_\tau^\star - z_2), \tag{23a}$$

$$\left\langle \log \nu^{(t+1)} - (1-\eta\tau)\log \nu^{(t)} - \eta\tau \log \nu_\tau^\star, z_2 - \nu_\tau^\star \right\rangle = -\eta(\nu_\tau^\star - z_1)^\top A(\nu_\tau^\star - z_2), \tag{23b}$$

and

$$\left\langle \log \zeta^{(t+1)} - (1-\eta\tau)\log \zeta^{(t)} - \eta\tau \log \zeta_\tau^\star, \zeta(z) - \zeta_\tau^\star \right\rangle = 0. \tag{24}$$

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 12 *For every $(\mu, \nu) \in \Delta(\mathcal{A}) \times \Delta(\mathcal{B})$, the following relations hold*

$$f_\tau(\mu_\tau^\star, \nu) - f_\tau(\mu, \nu_\tau^\star) = \tau \mathsf{KL}\big(\zeta \,\|\, \zeta_\tau^\star\big), \tag{25a}$$

$$f_\tau(\mu, \nu) - f_\tau(\mu_\tau^\star, \nu_\tau^\star) = (\mu_\tau^\star - \mu)^\top A(\nu_\tau^\star - \nu) + \tau \mathsf{KL}\big(\nu \,\|\, \nu_\tau^\star\big) - \tau \mathsf{KL}\big(\mu \,\|\, \mu_\tau^\star\big). \tag{25b}$$

In addition, we also make record of the following elementary lemma that is used frequently.

Lemma 13 *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

$$\|\log \mu_1 - \log \mu_2\|_\infty \le 2\|x_1 - x_2\|_\infty, \quad \text{and} \quad \mathsf{KL}\big(\mu_1 \,\|\, \mu_2\big) \le \frac{1}{2}\|x_1 - x_2 - c\cdot \mathbf{1}\|_\infty^2,$$

where the latter inequality holds for all $c \in \mathbb{R}$.

A.1 Proof of Proposition 1

Setting $\zeta(z) = \zeta^{(t+1)}$ in Lemma 11, we have

$$\big\langle \log \zeta^{(t+1)} - (1-\eta\tau)\log \zeta^{(t)} - \eta\tau \log \zeta_\tau^\star, \zeta^{(t+1)} - \zeta_\tau^\star \big\rangle = 0. \tag{26}$$

By the definition of the KL divergence, one has

$$\begin{aligned}
&- \big\langle \log \zeta^{(t+1)} - (1-\eta\tau)\log \zeta^{(t)} - \eta\tau \log \zeta_\tau^\star, \zeta_\tau^\star \big\rangle \\
&= -(1-\eta\tau)\big\langle \log \zeta_\tau^\star - \log \zeta^{(t)}, \zeta_\tau^\star \big\rangle + \big\langle \log \zeta_\tau^\star - \log \zeta^{(t+1)}, \zeta_\tau^\star \big\rangle \\
&= -(1-\eta\tau)\mathsf{KL}\big(\zeta_\tau^\star \,\|\, \zeta^{(t)}\big) + \mathsf{KL}\big(\zeta_\tau^\star \,\|\, \zeta^{(t+1)}\big),
\end{aligned} \tag{27}$$

and similarly,

$$\begin{aligned}
&\big\langle \log \zeta^{(t+1)} - (1-\eta\tau)\log \zeta^{(t)} - \eta\tau \log \zeta_\tau^\star, \zeta^{(t+1)} \big\rangle \\
&= (1-\eta\tau)\big\langle \log \zeta^{(t+1)} - \log \zeta^{(t)}, \zeta^{(t+1)} \big\rangle + \eta\tau \big\langle \log \zeta^{(t+1)} - \log \zeta_\tau^\star, \zeta^{(t+1)} \big\rangle \\
&= (1-\eta\tau)\mathsf{KL}\big(\zeta^{(t+1)} \,\|\, \zeta^{(t)}\big) + \eta\tau \mathsf{KL}\big(\zeta^{(t+1)} \,\|\, \zeta_\tau^\star\big).
\end{aligned}$$

Combining the above two equalities with (26), we arrive at

$$\mathsf{KL}\big(\zeta_\tau^\star \,\|\, \zeta^{(t+1)}\big) + \eta\tau \mathsf{KL}\big(\zeta^{(t+1)} \,\|\, \zeta_\tau^\star\big) + (1-\eta\tau)\mathsf{KL}\big(\zeta^{(t+1)} \,\|\, \zeta^{(t)}\big) = (1-\eta\tau)\mathsf{KL}\big(\zeta_\tau^\star \,\|\, \zeta^{(t)}\big). \tag{28}$$

This immediately leads to $\mathsf{KL}\big(\zeta_\tau^\star \,\|\, \zeta^{(t+1)}\big) \le (1-\eta\tau)\mathsf{KL}\big(\zeta_\tau^\star \,\|\, \zeta^{(t)}\big)$ by the nonnegativity of the KL divergence, as long as $1-\eta\tau \ge 0$. Therefore

$$\mathsf{KL}\big(\zeta_\tau^\star \,\|\, \zeta^{(t)}\big) \le (1-\eta\tau)^t \mathsf{KL}\big(\zeta_\tau^\star \,\|\, \zeta^{(0)}\big) \qquad \text{for all } t \ge 0.$$

A.2 Proof of Theorem 2

A.2.1 Proof of policy convergence in KL divergence (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 11 (cf. (24)) gives

$$\big\langle \log \zeta^{(t+1)} - (1-\eta\tau)\log \zeta^{(t)} - \eta\tau \log \zeta_\tau^\star, \bar{\zeta}^{(t+1)} - \zeta_\tau^\star \big\rangle = 0. \tag{29}$$

In view of the similarity of (26) and (29), 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 (27), we have

$$-\big\langle \log \zeta^{(t+1)} - (1-\eta\tau)\log \zeta^{(t)} - \eta\tau \log \zeta_\tau^\star, \zeta_\tau^\star \big\rangle = -(1-\eta\tau)\mathsf{KL}\big(\zeta_\tau^\star \,\|\, \zeta^{(t)}\big) + \mathsf{KL}\big(\zeta_\tau^\star \,\|\, \zeta^{(t+1)}\big). \tag{30}$$

On the other hand, it is easily seen that

$$\big\langle \log \zeta^{(t+1)} - (1-\eta\tau)\log \zeta^{(t)} - \eta\tau \log \zeta_\tau^\star, \bar{\zeta}^{(t+1)} \big\rangle$$

$$
\begin{aligned}
&= \big\langle \log \bar{\zeta}^{(t+1)} - (1-\eta\tau)\log \zeta^{(t)} - \eta\tau \log \zeta_\tau^\star, \bar{\zeta}^{(t+1)} \big\rangle - \big\langle \log \bar{\zeta}^{(t+1)} - \log \zeta^{(t+1)}, \zeta^{(t+1)} \big\rangle \\
&\quad - \big\langle \log \bar{\zeta}^{(t+1)} - \log \zeta^{(t+1)}, \bar{\zeta}^{(t+1)} - \zeta^{(t+1)} \big\rangle \\
&= (1-\eta\tau)\mathsf{KL}\big(\bar{\zeta}^{(t+1)} \,\|\, \zeta^{(t)}\big) + \eta\tau \mathsf{KL}\big(\bar{\zeta}^{(t+1)} \,\|\, \zeta_\tau^\star\big) + \mathsf{KL}\big(\zeta^{(t+1)} \,\|\, \bar{\zeta}^{(t+1)}\big) \\
&\quad - \big\langle \log \bar{\zeta}^{(t+1)} - \log \zeta^{(t+1)}, \bar{\zeta}^{(t+1)} - \zeta^{(t+1)} \big\rangle. \qquad (31)
\end{aligned}
$$

Combining equalities (30), (31) with (29), we are left with the following relation pertaining to bounding $\mathsf{KL}\big(\zeta_\tau^\star \,\|\, \zeta^{(t)}\big)$:

$$
\begin{aligned}
(1-\eta\tau)\mathsf{KL}\big(\zeta_\tau^\star \,\|\, \zeta^{(t)}\big) &= (1-\eta\tau)\mathsf{KL}\big(\bar{\zeta}^{(t+1)} \,\|\, \zeta^{(t)}\big) + \eta\tau \mathsf{KL}\big(\bar{\zeta}^{(t+1)} \,\|\, \zeta_\tau^\star\big) + \mathsf{KL}\big(\zeta^{(t+1)} \,\|\, \bar{\zeta}^{(t+1)}\big) \\
&\quad - \big\langle \log \bar{\zeta}^{(t+1)} - \log \zeta^{(t+1)}, \bar{\zeta}^{(t+1)} - \zeta^{(t+1)} \big\rangle + \mathsf{KL}\big(\zeta_\tau^\star \,\|\, \zeta^{(t+1)}\big). \qquad (32)
\end{aligned}
$$

In addition, to bound $\mathsf{KL}\big(\zeta_\tau^\star \,\|\, \bar{\zeta}^{(t+1)}\big)$, we will resort to the following three-point equality, which reads

$$
\begin{aligned}
&\mathsf{KL}\big(\zeta_\tau^\star \,\|\, \bar{\zeta}^{(t+1)}\big) \\
&= \mathsf{KL}\big(\zeta_\tau^\star \,\|\, \zeta^{(t+1)}\big) - \big\langle \zeta_\tau^\star, \log \bar{\zeta}^{(t+1)} - \log \zeta^{(t+1)} \big\rangle \\
&= \mathsf{KL}\big(\zeta_\tau^\star \,\|\, \zeta^{(t+1)}\big) - \mathsf{KL}\big(\bar{\zeta}^{(t+1)} \,\|\, \zeta^{(t+1)}\big) - \big\langle \zeta_\tau^\star - \bar{\zeta}^{(t+1)}, \log \bar{\zeta}^{(t+1)} - \log \zeta^{(t+1)} \big\rangle, \qquad (33)
\end{aligned}
$$

which can be checked directly using the definition of the KL divergence.

To proceed, we need to control $\big\langle \log \bar{\zeta}^{(t+1)} - \log \zeta^{(t+1)}, \bar{\zeta}^{(t+1)} - \zeta^{(t+1)} \big\rangle$ on the right-hand side of inequality (32), and $\big\langle \zeta_\tau^\star - \bar{\zeta}^{(t+1)}, \log \bar{\zeta}^{(t+1)} - \log \zeta^{(t+1)} \big\rangle$ on the right-hand side of inequality (33), 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

$$
\log \bar{\mu}^{(t+1)} - \log \mu^{(t+1)} = \eta A(\nu^{(t)} - \bar{\nu}^{(t+1)}) + c \cdot \mathbf{1} \qquad (34)
$$

for some normalization constant c. With this relation in place, one has

$$
\begin{aligned}
\big\langle \log \bar{\mu}^{(t+1)} - \log \mu^{(t+1)}, \bar{\mu}^{(t+1)} - \mu^{(t+1)} \big\rangle &= \eta(\bar{\mu}^{(t+1)} - \mu^{(t+1)})^\top A(\nu^{(t)} - \bar{\nu}^{(t+1)}) \\
&\le \eta \|A\|_\infty \left\|\bar{\mu}^{(t+1)} - \mu^{(t+1)}\right\|_1 \left\|\bar{\nu}^{(t+1)} - \nu^{(t)}\right\|_1.
\end{aligned}
$$

Combined with Pinsker's inequality, it is therefore clear that

$$
\begin{aligned}
\big\langle \log \bar{\mu}^{(t+1)} - \log \mu^{(t+1)}, \bar{\mu}^{(t+1)} - \mu^{(t+1)} \big\rangle &\le \frac{1}{2}\eta \|A\|_\infty \left(\left\|\bar{\mu}^{(t+1)} - \mu^{(t+1)}\right\|_1^2 + \left\|\bar{\nu}^{(t+1)} - \nu^{(t)}\right\|_1^2 \right) \\
&\le \eta \|A\|_\infty \Big(\mathsf{KL}\big(\mu^{(t+1)} \,\|\, \bar{\mu}^{(t+1)}\big) + \mathsf{KL}\big(\bar{\nu}^{(t+1)} \,\|\, \nu^{(t)}\big) \Big). \qquad (35)
\end{aligned}
$$

Analogously, one can achieve the same bound regarding the quantity

$$
\big\langle \log \bar{\nu}^{(t+1)} - \log \nu^{(t+1)}, \bar{\nu}^{(t+1)} - \nu^{(t+1)} \big\rangle.
$$

Summing up these two inequalities, we end up with

$$
\big\langle \log \bar{\zeta}^{(t+1)} - \log \zeta^{(t+1)}, \bar{\zeta}^{(t+1)} - \zeta^{(t+1)} \big\rangle \le \eta \|A\|_\infty \Big(\mathsf{KL}\big(\zeta^{(t+1)} \,\|\, \bar{\zeta}^{(t+1)}\big) + \mathsf{KL}\big(\bar{\zeta}^{(t+1)} \,\|\, \zeta^{(t)}\big) \Big).
$$

Plugging the above inequality into inequality (32) leads to

$$\begin{aligned}\mathsf{KL}\big(\zeta_\tau^\star \,\|\, \zeta^{(t+1)}\big) \le (1-\eta\tau)\mathsf{KL}\big(\zeta_\tau^\star \,\|\, \zeta^{(t)}\big) - \big(1-\eta\tau-\eta\,\|A\|_\infty\big)\mathsf{KL}\big(\bar\zeta^{(t+1)} \,\|\, \zeta^{(t)}\big) \\ - \eta\tau\mathsf{KL}\big(\bar\zeta^{(t+1)} \,\|\, \zeta_\tau^\star\big) - (1-\eta\,\|A\|_\infty)\mathsf{KL}\big(\zeta^{(t+1)} \,\|\, \bar\zeta^{(t+1)}\big). \end{aligned} \tag{36}$$

Therefore, as long as the learning rate η satisfies $\eta \le \frac{1}{\tau+\|A\|_\infty}$, we are ensured that

$$\mathsf{KL}\big(\zeta_\tau^\star \,\|\, \zeta^{(t+1)}\big) \le (1-\eta\tau)\mathsf{KL}\big(\zeta_\tau^\star \,\|\, \zeta^{(t)}\big),$$

which further implies inequality (9a) when applied recursively.

Bounding $\mathsf{KL}\big(\zeta_\tau^\star \,\|\, \bar\zeta^{(t+1)}\big)$ for PU. By similar tricks of arriving at (35), we have

$$\begin{aligned}-\big\langle \mu_\tau^\star - \bar\mu^{(t+1)}, \log\bar\mu^{(t+1)} - \log\mu^{(t+1)}\big\rangle &= -\eta(\mu_\tau^\star - \bar\mu^{(t+1)})^\top A(\nu^{(t)} - \bar\nu^{(t+1)}) \\ &\le \frac{1}{2}\eta\,\|A\|_\infty\left(\left\|\mu_\tau^\star - \bar\mu^{(t+1)}\right\|_1^2 + \left\|\nu^{(t)} - \bar\nu^{(t+1)}\right\|_1^2\right) \\ &\le \eta\,\|A\|_\infty\left(\mathsf{KL}\big(\mu_\tau^\star \,\|\, \bar\mu^{(t+1)}\big) + \mathsf{KL}\big(\bar\nu^{(t+1)} \,\|\, \nu^{(t)}\big)\right),\end{aligned}$$

following from (34) and Pinsker's inequality. A similar inequality for

$$-\big\langle \nu_\tau^\star - \bar\nu^{(t+1)}, \log\bar\nu^{(t+1)} - \log\nu^{(t+1)}\big\rangle$$

can be obtained by symmetry, and summing together the two leads to

$$-\big\langle \zeta_\tau^\star - \bar\zeta^{(t+1)}, \log\bar\zeta^{(t+1)} - \log\zeta^{(t+1)}\big\rangle \le \eta\,\|A\|_\infty\left(\mathsf{KL}\big(\zeta_\tau^\star \,\|\, \bar\zeta^{(t+1)}\big) + \mathsf{KL}\big(\bar\zeta^{(t+1)} \,\|\, \zeta^{(t)}\big)\right).$$

Plugging the above inequality into (33) and rearranging terms, we reach at

$$(1-\eta\,\|A\|_\infty)\mathsf{KL}\big(\zeta_\tau^\star \,\|\, \bar\zeta^{(t+1)}\big) \le \mathsf{KL}\big(\zeta_\tau^\star \,\|\, \zeta^{(t+1)}\big) + \eta\,\|A\|_\infty\,\mathsf{KL}\big(\bar\zeta^{(t+1)} \,\|\, \zeta^{(t)}\big).$$

Along with (36), we have

$$\begin{aligned}(1-\eta\,\|A\|_\infty)\mathsf{KL}\big(\zeta_\tau^\star \,\|\, \bar\zeta^{(t+1)}\big) \le (1-\eta\tau)\mathsf{KL}\big(\zeta_\tau^\star \,\|\, \zeta^{(t)}\big) - (1-\eta\tau-2\eta\,\|A\|_\infty)\mathsf{KL}\big(\bar\zeta^{(t+1)} \,\|\, \zeta^{(t)}\big) \\ - \eta\tau\mathsf{KL}\big(\bar\zeta^{(t+1)} \,\|\, \zeta^\star\big) - (1-\eta\,\|A\|_\infty)\mathsf{KL}\big(\zeta^{(t+1)} \,\|\, \bar\zeta^{(t+1)}\big).\end{aligned} \tag{37}$$

Therefore, with $\eta \le 1/(\tau + 2\,\|A\|_\infty)$ we have

$$\mathsf{KL}\big(\zeta_\tau^\star \,\|\, \bar\zeta^{(t+1)}\big) \le 2\mathsf{KL}\big(\zeta_\tau^\star \,\|\, \zeta^{(t)}\big) \le 2(1-\eta\tau)^t\mathsf{KL}\big(\zeta_\tau^\star \,\|\, \zeta^{(0)}\big).$$

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

$$\begin{aligned}\log\bar\mu^{(t+1)} - \log\mu^{(t+1)} &= \eta A(\bar\nu^{(t)} - \bar\nu^{(t+1)}) + c\cdot\mathbf{1} \\ &= \eta A(\bar\nu^{(t)} - \nu^{(t)}) + \eta A(\nu^{(t)} - \bar\nu^{(t+1)}) + c\cdot\mathbf{1},\end{aligned} \tag{38}$$

where c is some normalization constant. Similar to the proof of relation (35), it can be easily demonstrated that

$$\begin{aligned}
&\big\langle \log \bar{\mu}^{(t+1)} - \log \mu^{(t+1)}, \bar{\mu}^{(t+1)} - \mu^{(t+1)} \big\rangle \\
&= \eta(\bar{\mu}^{(t+1)} - \mu^{(t+1)})^\top A(\bar{\nu}^{(t)} - \nu^{(t)}) + \eta(\bar{\mu}^{(t+1)} - \mu^{(t+1)})^\top A(\nu^{(t)} - \bar{\nu}^{(t+1)}) \\
&\le \eta \left\|A\right\|_\infty \Big(\mathsf{KL}\big(\nu^{(t)} \,\|\, \bar{\nu}^{(t)}\big) + \mathsf{KL}\big(\bar{\nu}^{(t+1)} \,\|\, \nu^{(t)}\big) + 2\mathsf{KL}\big(\mu^{(t+1)} \,\|\, \bar{\mu}^{(t+1)}\big) \Big). \qquad (39)
\end{aligned}$$

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

$$\begin{aligned}
&\big\langle \log \bar{\zeta}^{(t+1)} - \log \zeta^{(t+1)}, \bar{\zeta}^{(t+1)} - \zeta^{(t+1)} \big\rangle \\
&\le \eta \left\|A\right\|_\infty \Big(\mathsf{KL}\big(\zeta^{(t)} \,\|\, \bar{\zeta}^{(t)}\big) + \mathsf{KL}\big(\bar{\zeta}^{(t+1)} \,\|\, \zeta^{(t)}\big) + 2\mathsf{KL}\big(\zeta^{(t+1)} \,\|\, \bar{\zeta}^{(t+1)}\big) \Big).
\end{aligned}$$

Plugging the above inequality into equation (32) and re-organizing terms, we arrive at

$$\begin{aligned}
&\mathsf{KL}\big(\zeta_\tau^\star \,\|\, \zeta^{(t+1)}\big) \\
&\le (1-\eta\tau)\mathsf{KL}\big(\zeta_\tau^\star \,\|\, \zeta^{(t)}\big) - (1-\eta\tau-\eta\left\|A\right\|_\infty)\mathsf{KL}\big(\bar{\zeta}^{(t+1)} \,\|\, \zeta^{(t)}\big) - \eta\tau\mathsf{KL}\big(\bar{\zeta}^{(t+1)} \,\|\, \zeta_\tau^\star\big) \\
&- (1-2\eta\left\|A\right\|_\infty)\mathsf{KL}\big(\zeta^{(t+1)} \,\|\, \bar{\zeta}^{(t+1)}\big) + \eta\left\|A\right\|_\infty \mathsf{KL}\big(\zeta^{(t)} \,\|\, \bar{\zeta}^{(t)}\big). \qquad (40)
\end{aligned}$$

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 (40) gives

$$\begin{aligned}
&\mathsf{KL}\big(\zeta_\tau^\star \,\|\, \zeta^{(t+1)}\big) + (1-2\eta\left\|A\right\|_\infty)\mathsf{KL}\big(\zeta^{(t+1)} \,\|\, \bar{\zeta}^{(t+1)}\big) \\
&\le (1-\eta\tau)\mathsf{KL}\big(\zeta_\tau^\star \,\|\, \zeta^{(t)}\big) + \eta\left\|A\right\|_\infty \mathsf{KL}\big(\zeta^{(t)} \,\|\, \bar{\zeta}^{(t)}\big) - \eta\tau\mathsf{KL}\big(\bar{\zeta}^{(t+1)} \,\|\, \zeta_\tau^\star\big). \\
&\le (1-\eta\tau)\Big[\mathsf{KL}\big(\zeta_\tau^\star \,\|\, \zeta^{(t)}\big) + (1-2\eta\left\|A\right\|_\infty)\mathsf{KL}\big(\zeta^{(t)} \,\|\, \bar{\zeta}^{(t)}\big)\Big] - \eta\tau\mathsf{KL}\big(\bar{\zeta}^{(t+1)} \,\|\, \zeta_\tau^\star\big).
\end{aligned}$$

For conciseness, let us introduce the shorthand notation

$$L^{(t)} := \mathsf{KL}\big(\zeta_\tau^\star \,\|\, \zeta^{(t)}\big) + (1-2\eta\left\|A\right\|_\infty)\mathsf{KL}\big(\zeta^{(t)} \,\|\, \bar{\zeta}^{(t)}\big). \qquad (41)$$

As a result, the above inequality can be restated as

$$L^{(t+1)} \le (1-\eta\tau)L^{(t)} - \eta\tau\mathsf{KL}\big(\bar{\zeta}^{(t+1)} \,\|\, \zeta_\tau^\star\big). \qquad (42)$$

Since we initialize OMWU with $\bar{\zeta}^{(0)} = \zeta^{(0)}$, therefore $L^{(0)} = \mathsf{KL}\big(\zeta_\tau^\star \,\|\, \zeta^{(0)}\big)$, which in turn gives

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

We complete the proof of inequality (9a) for OMWU.

Bounding $\mathsf{KL}\big(\zeta_\tau^\star \,\|\, \bar{\zeta}^{(t+1)}\big)$ for OMWU. By similar tricks of arriving at (39), we have

$$\begin{aligned}
&-\big\langle \mu_\tau^\star - \bar{\mu}^{(t+1)}, \log \bar{\mu}^{(t+1)} - \log \mu^{(t+1)} \big\rangle \\
&= \eta(\bar{\mu}^{(t+1)} - \mu_\tau^\star)^\top A(\bar{\nu}^{(t)} - \nu^{(t)}) + \eta(\bar{\mu}^{(t+1)} - \mu_\tau^\star)^\top A(\nu^{(t)} - \bar{\nu}^{(t+1)}) \\
&\le \eta \|A\|_\infty \Big(\mathsf{KL}\big(\nu^{(t)} \,\|\, \bar{\nu}^{(t)}\big) + \mathsf{KL}\big(\bar{\nu}^{(t+1)} \,\|\, \nu^{(t)}\big) + 2\mathsf{KL}\big(\mu_\tau^\star \,\|\, \bar{\mu}^{(t+1)}\big)\Big),
\end{aligned}$$

where the first line follows from (38). A similar inequality also holds for

$$-\big\langle \nu_\tau^\star - \bar{\nu}^{(t+1)}, \log \bar{\nu}^{(t+1)} - \log \nu^{(t+1)} \big\rangle.$$

Summing the two inequalities leads to

$$\begin{aligned}
&-\big\langle \zeta_\tau^\star - \bar{\zeta}^{(t+1)}, \log \bar{\zeta}^{(t+1)} - \log \zeta^{(t+1)} \big\rangle \\
&\le \eta \|A\|_\infty \Big(\mathsf{KL}\big(\zeta^{(t)} \,\|\, \bar{\zeta}^{(t)}\big) + \mathsf{KL}\big(\bar{\zeta}^{(t+1)} \,\|\, \zeta^{(t)}\big) + 2\mathsf{KL}\big(\zeta_\tau^\star \,\|\, \bar{\zeta}^{(t+1)}\big)\Big).
\end{aligned}$$

Plugging the above inequality into (33) and rearranging terms, we reach at

$$(1-2\eta\|A\|_\infty)\mathsf{KL}\big(\zeta_\tau^\star \,\|\, \bar{\zeta}^{(t+1)}\big) \le \mathsf{KL}\big(\zeta_\tau^\star \,\|\, \zeta^{(t+1)}\big) + \eta\|A\|_\infty \Big(\mathsf{KL}\big(\zeta^{(t)} \,\|\, \bar{\zeta}^{(t)}\big) + \mathsf{KL}\big(\bar{\zeta}^{(t+1)} \,\|\, \zeta^{(t)}\big)\Big).$$

Along with (40), we have

$$\begin{aligned}
&(1-2\eta\|A\|_\infty)\mathsf{KL}\big(\zeta_\tau^\star \,\|\, \bar{\zeta}^{(t+1)}\big) \\
&\le (1-\eta\tau)\mathsf{KL}\big(\zeta_\tau^\star \,\|\, \zeta^{(t)}\big) - (1-\eta\tau-2\eta\|A\|_\infty)\mathsf{KL}\big(\bar{\zeta}^{(t+1)} \,\|\, \zeta^{(t)}\big) - \eta\tau\mathsf{KL}\big(\bar{\zeta}^{(t+1)} \,\|\, \zeta_\tau^\star\big) \\
&\qquad - (1-2\eta\|A\|_\infty)\mathsf{KL}\big(\zeta^{(t+1)} \,\|\, \bar{\zeta}^{(t+1)}\big) + 2\eta\|A\|_\infty \mathsf{KL}\big(\zeta^{(t)} \,\|\, \bar{\zeta}^{(t)}\big) \\
&\le (1-\eta\tau)\mathsf{KL}\big(\zeta_\tau^\star \,\|\, \zeta^{(t)}\big) + 2\eta\|A\|_\infty \mathsf{KL}\big(\zeta^{(t)} \,\|\, \bar{\zeta}^{(t)}\big) \\
&\le \mathsf{KL}\big(\zeta_\tau^\star \,\|\, \zeta^{(t)}\big) + (1-2\eta\|A\|_\infty)\mathsf{KL}\big(\zeta^{(t)} \,\|\, \bar{\zeta}^{(t)}\big) =: L^{(t)},
\end{aligned}$$

where we recall the shorthand notation $L^{(t)}$ in (41). As the learning rate of OMWU satisfies $0 < \eta < \min\Big\{\frac{1}{2\|A\|_\infty + 2\tau},\ \frac{1}{4\|A\|_\infty}\Big\}$, it is clear that

$$\mathsf{KL}\big(\zeta_\tau^\star \,\|\, \bar{\zeta}^{(t+1)}\big) \le 2L^{(t)} \overset{\text{(i)}}{\le} 2(1-\eta\tau)^t L^{(0)} \le 2(1-\eta\tau)^t \mathsf{KL}\big(\zeta_\tau^\star \,\|\, \zeta^{(0)}\big),$$

where (i) follows from the recursive relation $L^{(t+1)} \le (1-\eta\tau)L^{(t)}$ shown in inequality (42).

A.2.2 Proof of entrywise convergence of policy log-ratios (9b)

To facilitate the proof, we introduce an auxiliary sequence $\{\xi^{(t)} \in \mathbb{R}^{|\mathcal{A}|}\}$ constructed recursively by

$$\xi^{(0)}(a) = \|\exp(A\nu_\tau^\star/\tau)\|_1 \cdot \mu^{(0)}(a), \tag{43a}$$

$$\xi^{(t+1)}(a) = \xi^{(t)}(a)^{1-\eta\tau} \exp(\eta[A\bar{\nu}^{(t+1)}]_a), \qquad \forall a \in \mathcal{A}, t \ge 0. \tag{43b}$$

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

$$\Big\|\log \mu^{(t+1)} - \log \mu_\tau^\star\Big\|_\infty \le 2\Big\|\log \xi^{(t+1)} - A\nu_\tau^\star/\tau\Big\|_\infty, \tag{44}$$

where we make use of Lemma 13.

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 (44). Taking logarithm on both sides of (43b) 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

$$\begin{aligned}\left\|\log \xi^{(t+1)} - A\nu_\tau^\star/\tau\right\|_\infty &\le (1-\eta\tau)\left\|\log \xi^{(t)} - A\nu_\tau^\star/\tau\right\|_\infty + \eta\left\|A\right\|_\infty \left\|\bar{\nu}^{(t+1)} - \nu_\tau^\star\right\|_1 \\ &\le (1-\eta\tau)\left\|\log \xi^{(t)} - A\nu_\tau^\star/\tau\right\|_\infty + \eta\left\|A\right\|_\infty \left[2\mathsf{KL}\big(\nu_\tau^\star \,\|\, \bar{\nu}^{(t+1)}\big)\right]^{1/2} \\ &\le (1-\eta\tau)\left\|\log \xi^{(t)} - A\nu_\tau^\star/\tau\right\|_\infty + \eta\left\|A\right\|_\infty \left[2\mathsf{KL}\big(\zeta_\tau^\star \,\|\, \bar{\zeta}^{(t+1)}\big)\right]^{1/2}.\end{aligned} \tag{45}$$

Plugging the bound of $\mathsf{KL}\big(\zeta_\tau^\star \,\|\, \bar{\zeta}^{(t+1)}\big)$ from relation (9a) into (45) and invoking the inequality recursively leads to

$$\begin{aligned}&\left\|\log \xi^{(t+1)} - A\nu_\tau^\star/\tau\right\|_\infty \\ &\le (1-\eta\tau)^{t+1}\left\|\log \xi^{(0)} - A\nu_\tau^\star/\tau\right\|_\infty + 2\eta\left\|A\right\|_\infty \sum_{s=1}^{t+1}(1-\eta\tau)^{t+1-s/2}\mathsf{KL}\big(\zeta_\tau^\star \,\|\, \zeta^{(0)}\big)^{1/2} \\ &\le (1-\eta\tau)^{t+1}\left\|\log \xi^{(0)} - A\nu_\tau^\star/\tau\right\|_\infty + 2\eta\left\|A\right\|_\infty (1-\eta\tau)^{(t+1)/2}\frac{1}{1-(1-\eta\tau)^{1/2}}\mathsf{KL}\big(\zeta_\tau^\star \,\|\, \zeta^{(0)}\big)^{1/2} \\ &\le (1-\eta\tau)^{t+1}\left\|\log \xi^{(0)} - A\nu_\tau^\star/\tau\right\|_\infty + 4\tau^{-1}\left\|A\right\|_\infty (1-\eta\tau)^{(t+1)/2}\mathsf{KL}\big(\zeta_\tau^\star \,\|\, \zeta^{(0)}\big)^{1/2},\end{aligned}$$

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

$$\begin{aligned}&\left\|\log \mu^{(t+1)} - \log \mu_\tau^\star\right\|_\infty \\ &\le 2\left\|\log \xi^{(t+1)} - A\nu_\tau^\star/\tau\right\|_\infty \\ &\le 2(1-\eta\tau)^{t+1}\left\|\log \xi^{(0)} - A\nu_\tau^\star/\tau\right\|_\infty + 8\tau^{-1}\left\|A\right\|_\infty (1-\eta\tau)^{(t+1)/2}\mathsf{KL}\big(\zeta_\tau^\star \,\|\, \zeta^{(0)}\big)^{1/2} \\ &\le 2(1-\eta\tau)^{t+1}\left\|\log \mu^{(0)} - \log \mu_\tau^\star\right\|_\infty + 8\tau^{-1}\left\|A\right\|_\infty (1-\eta\tau)^{(t+1)/2}\mathsf{KL}\big(\zeta_\tau^\star \,\|\, \zeta^{(0)}\big)^{1/2}.\end{aligned}$$

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 (9b).

A.2.3 Proof of convergence of optimality gap (9c)

To streamline our discussions, we only provide the proof of inequality (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 11 inequality (23a) gives

$$\big\langle \log \mu^{(t+1)} - (1-\eta\tau)\log \mu^{(t)} - \eta\tau \log \mu_\tau^\star, \bar{\mu}^{(t+1)} - \mu_\tau^\star \big\rangle = \eta(\mu_\tau^\star - \bar{\mu}^{(t+1)})^\top A(\nu_\tau^\star - \bar{\nu}^{(t+1)}). \tag{46}$$

Similar to what we have done in the proof of (9a) (cf. (32)), based on the above relation, we can therefore rearrange terms and conclude that

$$\begin{aligned}
&\eta\Big(\tau \mathsf{KL}\big(\bar{\mu}^{(t+1)} \,\|\, \mu_\tau^\star\big) - (\mu_\tau^\star - \bar{\mu}^{(t+1)})^\top A(\nu_\tau^\star - \bar{\nu}^{(t+1)})\Big) \\
&= (1-\eta\tau)\mathsf{KL}\big(\mu_\tau^\star \,\|\, \mu^{(t)}\big) - (1-\eta\tau)\mathsf{KL}\big(\bar{\mu}^{(t+1)} \,\|\, \mu^{(t)}\big) - \mathsf{KL}\big(\mu^{(t+1)} \,\|\, \bar{\mu}^{(t+1)}\big) \\
&\quad + \big\langle \log \bar{\mu}^{(t+1)} - \log \mu^{(t+1)}, \bar{\mu}^{(t+1)} - \mu^{(t+1)} \big\rangle - \mathsf{KL}\big(\mu_\tau^\star \,\|\, \mu^{(t+1)}\big).
\end{aligned} \tag{47}$$

In conjunction with Lemma 12 (cf. (25b)), we can further derive

$$\begin{aligned}
&\eta\big(f_\tau(\mu_\tau^\star, \nu_\tau^\star) - f_\tau(\bar{\mu}^{(t+1)}, \bar{\nu}^{(t+1)})\big) \le \eta\Big(\tau \mathsf{KL}\big(\bar{\mu}^{(t+1)} \,\|\, \mu_\tau^\star\big) - (\mu_\tau^\star - \bar{\mu}^{(t+1)})^\top A(\nu_\tau^\star - \bar{\nu}^{(t+1)})\Big) \\
&= (1-\eta\tau)\mathsf{KL}\big(\mu_\tau^\star \,\|\, \mu^{(t)}\big) - (1-\eta\tau)\mathsf{KL}\big(\bar{\mu}^{(t+1)} \,\|\, \mu^{(t)}\big) - \mathsf{KL}\big(\mu_\tau^\star \,\|\, \mu^{(t+1)}\big) \\
&\quad - \mathsf{KL}\big(\mu^{(t+1)} \,\|\, \bar{\mu}^{(t+1)}\big) + \big\langle \log \bar{\mu}^{(t+1)} - \log \mu^{(t+1)}, \bar{\mu}^{(t+1)} - \mu^{(t+1)} \big\rangle,
\end{aligned} \tag{48}$$

where the second line follows from (47). From this point, we shall continue the proofs for PU and OMWU separately but follow similar strategies.

Remaining steps for PU. Plugging relation (35) into (48), we arrive at

$$\begin{aligned}
&\eta\big(f_\tau(\mu_\tau^\star, \nu_\tau^\star) - f_\tau(\bar{\mu}^{(t+1)}, \bar{\nu}^{(t+1)})\big) \\
&\le (1-\eta\tau)\mathsf{KL}\big(\mu_\tau^\star \,\|\, \mu^{(t)}\big) - (1-\eta\tau)\mathsf{KL}\big(\bar{\mu}^{(t+1)} \,\|\, \mu^{(t)}\big) - \mathsf{KL}\big(\mu_\tau^\star \,\|\, \mu^{(t+1)}\big) \\
&\quad - (1-\eta\|A\|_\infty)\mathsf{KL}\big(\mu^{(t+1)} \,\|\, \bar{\mu}^{(t+1)}\big) + \eta\|A\|_\infty \mathsf{KL}\big(\bar{\nu}^{(t+1)} \,\|\, \nu^{(t)}\big) \\
&\le (1-\eta\tau)\mathsf{KL}\big(\mu_\tau^\star \,\|\, \mu^{(t)}\big) - \mathsf{KL}\big(\mu_\tau^\star \,\|\, \mu^{(t+1)}\big) - (1-\eta\tau)\mathsf{KL}\big(\bar{\mu}^{(t+1)} \,\|\, \mu^{(t)}\big) \\
&\quad + \eta\|A\|_\infty \mathsf{KL}\big(\bar{\nu}^{(t+1)} \,\|\, \nu^{(t)}\big),
\end{aligned} \tag{49}$$

where the last line holds since $\eta(\tau + \|A\|_\infty) \le 1$. Similarly, from Lemma 11 inequality (23b), one can establish the following inequality in parallel

$$\begin{aligned}
&\eta\big(f_\tau(\bar{\mu}^{(t+1)}, \bar{\nu}^{(t+1)}) - f_\tau(\mu_\tau^\star, \nu_\tau^\star)\big) \\
&\le (1-\eta\tau)\mathsf{KL}\big(\nu_\tau^\star \,\|\, \nu^{(t)}\big) - \mathsf{KL}\big(\nu_\tau^\star \,\|\, \nu^{(t+1)}\big) - (1-\eta\tau)\mathsf{KL}\big(\bar{\nu}^{(t+1)} \,\|\, \nu^{(t)}\big) \\
&\quad + \eta\|A\|_\infty \mathsf{KL}\big(\bar{\mu}^{(t+1)} \,\|\, \mu^{(t)}\big).
\end{aligned} \tag{50}$$

We are ready to establish inequality (9c) for PU. Computing (49) $+\frac{2}{3}\cdot$ (50) gives

$$\begin{aligned}
&\frac{\eta}{3}\big(f_\tau(\mu_\tau^\star, \nu_\tau^\star) - f_\tau(\bar{\mu}^{(t+1)}, \bar{\nu}^{(t+1)})\big) \\
&\le (1-\eta\tau)\left[\mathsf{KL}\big(\mu_\tau^\star \,\|\, \mu^{(t)}\big) + \frac{2}{3}\mathsf{KL}\big(\nu_\tau^\star \,\|\, \nu^{(t)}\big)\right] - \left[\mathsf{KL}\big(\mu_\tau^\star \,\|\, \mu^{(t+1)}\big) + \frac{2}{3}\mathsf{KL}\big(\nu_\tau^\star \,\|\, \nu^{(t+1)}\big)\right]
\end{aligned}$$

$$
\begin{aligned}
&- \left[(1-\eta\tau) - \frac{2}{3}\eta\left\|A\right\|_\infty\right] \mathsf{KL}\big(\bar{\mu}^{(t+1)} \,\|\, \mu^{(t)}\big) + \left[\eta\left\|A\right\|_\infty - \frac{2}{3}(1-\eta\tau)\right] \mathsf{KL}\big(\bar{\nu}^{(t+1)} \,\|\, \nu^{(t)}\big) \\
&\le (1-\eta\tau)\left[\mathsf{KL}\big(\mu_\tau^\star \,\|\, \mu^{(t)}\big) + \frac{2}{3}\mathsf{KL}\big(\nu_\tau^\star \,\|\, \nu^{(t)}\big)\right] - \left[\mathsf{KL}\big(\mu_\tau^\star \,\|\, \mu^{(t+1)}\big) + \frac{2}{3}\mathsf{KL}\big(\nu_\tau^\star \,\|\, \nu^{(t+1)}\big)\right] \quad (51)
\end{aligned}
$$

Here, the last step is due to the fact that $(1-\eta\tau) - \frac{2}{3}\eta\left\|A\right\|_\infty \ge 0$ and $\eta\left\|A\right\|_\infty - \frac{2}{3}(1-\eta\tau) \le 0$ when $0 < \eta \le \frac{1}{\tau+2\|A\|_\infty}$. As a direct consequence, the difference $f_\tau(\mu_\tau^\star, \nu_\tau^\star) - f_\tau(\bar{\mu}^{(t)}, \bar{\nu}^{(t)})$ satisfies

$$
\begin{aligned}
&\frac{\eta}{3}\big(f_\tau(\mu_\tau^\star, \nu_\tau^\star) - f_\tau(\bar{\mu}^{(t)}, \bar{\nu}^{(t)})\big) \\
&\le (1-\eta\tau)\left[(1-\eta\tau)\mathsf{KL}\big(\mu_\tau^\star \,\|\, \mu^{(t-1)}\big) + \eta\left\|A\right\|_\infty \mathsf{KL}\big(\nu_\tau^\star \,\|\, \nu^{(t-1)}\big)\right] \\
&\le (1-\eta\tau)\mathsf{KL}\big(\zeta_\tau^\star \,\|\, \zeta^{(t-1)}\big) \le (1-\eta\tau)^t \mathsf{KL}\big(\zeta_\tau^\star \,\|\, \zeta^{(0)}\big).
\end{aligned}
$$

We conclude by noting that the other side of (9c) can be shown by considering $\frac{2}{3}\cdot$ (49) + (50) combined with similar arguments, and are therefore omitted.

Remaining steps for OMWU. Similar to the case of PU, plugging (39) into (48) gives

$$
\begin{aligned}
&\eta\big(f_\tau(\mu_\tau^\star, \nu_\tau^\star) - f_\tau(\bar{\mu}^{(t+1)}, \bar{\nu}^{(t+1)})\big) \\
&\le (1-\eta\tau)\mathsf{KL}\big(\mu_\tau^\star \,\|\, \mu^{(t)}\big) - (1-\eta\tau)\mathsf{KL}\big(\bar{\mu}^{(t+1)} \,\|\, \mu^{(t)}\big) - \mathsf{KL}\big(\mu_\tau^\star \,\|\, \mu^{(t+1)}\big) \\
&\quad - (1-2\eta\left\|A\right\|_\infty)\mathsf{KL}\big(\mu^{(t+1)} \,\|\, \bar{\mu}^{(t+1)}\big) + \eta\left\|A\right\|_\infty\left[\mathsf{KL}\big(\nu^{(t)} \,\|\, \bar{\nu}^{(t)}\big) + \mathsf{KL}\big(\bar{\nu}^{(t+1)} \,\|\, \nu^{(t)}\big)\right].
\end{aligned} \quad (52)
$$

Similarly, one can establish a symmetric inequality as follows

$$
\begin{aligned}
&\eta\big(f_\tau(\bar{\mu}^{(t+1)}, \bar{\nu}^{(t+1)}) - f_\tau(\mu_\tau^\star, \nu_\tau^\star)\big) \\
&\le (1-\eta\tau)\mathsf{KL}\big(\nu_\tau^\star \,\|\, \nu^{(t)}\big) - (1-\eta\tau)\mathsf{KL}\big(\bar{\nu}^{(t+1)} \,\|\, \nu^{(t)}\big) - \mathsf{KL}\big(\nu_\tau^\star \,\|\, \nu^{(t+1)}\big) \\
&\quad - (1-2\eta\left\|A\right\|_\infty)\mathsf{KL}\big(\nu^{(t+1)} \,\|\, \bar{\nu}^{(t+1)}\big) + \eta\left\|A\right\|_\infty\left[\mathsf{KL}\big(\mu^{(t)} \,\|\, \bar{\mu}^{(t)}\big) + \mathsf{KL}\big(\bar{\mu}^{(t+1)} \,\|\, \mu^{(t)}\big)\right].
\end{aligned} \quad (53)
$$

Directly computing (52) $+\frac{2}{3}\cdot$ (53) gives

$$
\begin{aligned}
&\frac{\eta}{3}\cdot\big(f_\tau(\mu_\tau^\star, \nu_\tau^\star) - f_\tau(\bar{\mu}^{(t+1)}, \bar{\nu}^{(t+1)})\big) \\
&\le (1-\eta\tau)\left[\mathsf{KL}\big(\mu_\tau^\star \,\|\, \mu^{(t)}\big) + \frac{2}{3}\mathsf{KL}\big(\nu_\tau^\star \,\|\, \nu^{(t)}\big)\right] - \left[\mathsf{KL}\big(\mu_\tau^\star \,\|\, \mu^{(t+1)}\big) + \frac{2}{3}\mathsf{KL}\big(\nu_\tau^\star \,\|\, \nu^{(t+1)}\big)\right] \\
&\quad - \left[(1-\eta\tau) - \frac{2}{3}\eta\left\|A\right\|_\infty\right] \mathsf{KL}\big(\bar{\mu}^{(t+1)} \,\|\, \mu^{(t)}\big) + \left[\eta\left\|A\right\|_\infty - \frac{2}{3}(1-\eta\tau)\right] \mathsf{KL}\big(\bar{\nu}^{(t+1)} \,\|\, \nu^{(t)}\big) \\
&\quad + \eta\left\|A\right\|_\infty\left[\frac{2}{3}\mathsf{KL}\big(\mu^{(t)} \,\|\, \bar{\mu}^{(t)}\big) + \mathsf{KL}\big(\nu^{(t)} \,\|\, \bar{\nu}^{(t)}\big)\right] \\
&\quad - (1-2\eta\left\|A\right\|_\infty)\left[\mathsf{KL}\big(\mu^{(t+1)} \,\|\, \bar{\mu}^{(t+1)}\big) + \frac{2}{3}\mathsf{KL}\big(\nu^{(t+1)} \,\|\, \bar{\nu}^{(t+1)}\big)\right].
\end{aligned} \quad (54)
$$

With our choice of the learning rate $\eta \le \min\{\frac{1}{2\|A\|_\infty+2\tau}, \frac{1}{4\|A\|_\infty}\}$, it is guarantees that

$$\eta\|A\|_\infty - \frac{2}{3}(1-\eta\tau) \le 0, \quad (1-\eta\tau) - \frac{2}{3}\eta\|A\|_\infty \ge 0 \quad \text{and} \quad (1-\eta\tau)(1-2\eta\|A\|_\infty) \ge \frac{3}{2}\eta\|A\|_\infty.$$

To proceed, let us introduce the shorthand notation

$$\begin{aligned} G^{(t)} := \mathsf{KL}\big(\mu_\tau^\star \,\|\, \mu^{(t)}\big) + \frac{2}{3}\mathsf{KL}\big(\nu_\tau^\star \,\|\, \nu^{(t)}\big) \\ + \frac{2}{3}(1-2\eta\|A\|_\infty)\Big[\mathsf{KL}\big(\mu^{(t)} \,\|\, \bar{\mu}^{(t)}\big) + \mathsf{KL}\big(\nu^{(t)} \,\|\, \bar{\nu}^{(t)}\big)\Big]. \end{aligned}$$

With this piece of notation, we can write inequality (54) as

$$\frac{\eta}{3}(f_\tau(\mu_\tau^\star, \nu_\tau^\star) - f_\tau(\bar{\mu}^{(t+1)}, \bar{\nu}^{(t+1)})) \le (1-\eta\tau)G^{(t)} - G^{(t+1)}, \tag{55}$$

which in turn implies

$$\begin{aligned} &\frac{\eta}{3}(f_\tau(\mu_\tau^\star, \nu_\tau^\star) - f_\tau(\bar{\mu}^{(t)}, \bar{\nu}^{(t)})) \\ &\le (1-\eta\tau)G^{(t-1)} \le (1-\eta\tau)L^{(t-1)} \le (1-\eta\tau)^t L^{(0)} = (1-\eta\tau)^t \mathsf{KL}\big(\zeta_\tau^\star \,\|\, \zeta^{(0)}\big), \end{aligned}$$

with $L^{(t)}$ defined in (41). This finishes the proof of (9c) for OMWU.

A.2.4 Proof of convergence of duality gap (9d)

The proof of inequality (9d) is built upon the following lemma whose proof is deferred to Appendix C.4.

Lemma 14 *The duality gap at $\zeta = (\mu, \nu)$ can be bounded as*

$$\max_{\mu' \in \Delta(\mathcal{A})} f_\tau(\mu', \nu) - \min_{\nu' \in \Delta(\mathcal{B})} f_\tau(\mu, \nu') \le \tau\mathsf{KL}\big(\zeta \,\|\, \zeta_\tau^\star\big) + \tau^{-1}\|A\|_\infty^2 \mathsf{KL}\big(\zeta_\tau^\star \,\|\, \zeta\big).$$

Applying Lemma 14 to $\bar{\zeta}^{(t)} = (\bar{\mu}^{(t)}, \bar{\nu}^{(t)})$ yields

$$\begin{aligned} \mathsf{DualGap}_\tau(\bar{\zeta}^{(t)}) &\le \tau\mathsf{KL}\big(\bar{\zeta}^{(t)} \,\|\, \zeta_\tau^\star\big) + \tau^{-1}\|A\|_\infty^2 \mathsf{KL}\big(\zeta_\tau^\star \,\|\, \bar{\zeta}^{(t)}\big) \\ &\le \tau\mathsf{KL}\big(\bar{\zeta}^{(t)} \,\|\, \zeta_\tau^\star\big) + 2\tau^{-1}\|A\|_\infty^2 (1-\eta\tau)^{t-1} \mathsf{KL}\big(\zeta_\tau^\star \,\|\, \zeta^{(0)}\big), \end{aligned} \tag{56}$$

where the second step results from (9a). It remains to bound $\tau\mathsf{KL}\big(\bar{\zeta}^{(t)} \,\|\, \zeta_\tau^\star\big)$, which we proceed separately for PU and OMWU.

Remaining steps for PU. From inequality (36), we are ensured that

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

It thus follows that

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

where the last inequality is due to inequality (9a). Plugging the above inequality into (56) completes the proof of inequality (9d) for PU.

Remaining steps for OMWU. From inequality (42), we are ensured that

$$\tau \mathsf{KL}\big(\bar{\zeta}^{(t)} \,\|\, \zeta_\tau^\star\big) \le \eta^{-1}(1-\eta\tau)L^{(t-1)} \le \eta^{-1}(1-\eta\tau)^t L^{(0)} = \eta^{-1}(1-\eta\tau)^t \mathsf{KL}\big(\zeta_\tau^\star \,\|\, \zeta^{(0)}\big),$$

where the last equality follows from $L^{(0)} = \mathsf{KL}\big(\zeta_\tau^\star \,\|\, \zeta^{(0)}\big)$. Plugging the above inequality into (56) finishes the proof of inequality (9d) for OMWU.

A.3 Proof of Theorem 7

To begin, we note that in the no-regret setting, upon receiving $A^\top \bar{\mu}^{(t)}$ (which is possibly adversarial), the update rule of player 2 is given by

$$\nu^{(t)}(b) \propto \nu^{(t-1)}(b)^{1-\eta_{t-1}\tau} \exp(-\eta_{t-1}[A^\top \bar{\mu}^{(t)}]_b), \tag{57a}$$

$$\bar{\nu}^{(t+1)}(b) \propto \nu^{(t)}(b)^{1-\eta_t\tau} \exp(-\eta_t[A^\top \bar{\mu}^{(t)}]_b). \tag{57b}$$

Recalling $f_\tau^{(t)}(\nu) = \bar{\mu}^{(t)\top} A\nu + \tau\mathcal{H}(\bar{\mu}^{(t)}) - \tau\mathcal{H}(\nu)$, we introduce an important quantity which is the gradient of $f_\tau^{(t)}(\nu)$ at $\bar{\nu}^{(t)}$:

$$\overline{\nabla}^{(t)} := \nabla_\nu f_\tau^{(t)}(\nu)\Big|_{\nu=\bar{\nu}^{(t)}} = A^\top \bar{\mu}^{(t)} + \tau(\log \bar{\nu}^{(t)} + \mathbf{1}). \tag{58}$$

The following lemma presents an ℓ_∞ bound on the size of $\overline{\nabla}^{(t)}$, whose proof can be found in Appendix C.5.

Lemma 15 *It holds for all $t \ge 0$ that*

$$\left\|\overline{\nabla}^{(t)}\right\|_\infty \le \tau \log|\mathcal{B}| + 3\|A\|_\infty.$$

Regret decomposition. By the definition of $f_\tau^{(t)}(\nu)$, we have

$$\begin{aligned} f_\tau^{(t)}(\bar{\nu}^{(t)}) - f_\tau^{(t)}(\nu) &= \langle A^\top \bar{\mu}^{(t)}, \bar{\nu}^{(t)} - \nu\rangle + \tau\bar{\nu}^{(t)\top}\log\bar{\nu}^{(t)} - \tau\nu^\top \log\nu \\ &= (A^\top\bar{\mu}^{(t)} + \tau\log\bar{\nu}^{(t)})^\top(\bar{\nu}^{(t)} - \nu) - \tau\mathsf{KL}\big(\nu \,\|\, \bar{\nu}^{(t)}\big) \\ &= \langle\overline{\nabla}^{(t)}, \bar{\nu}^{(t)} - \nu\rangle - \tau\mathsf{KL}\big(\nu \,\|\, \bar{\nu}^{(t)}\big), \end{aligned} \tag{59}$$

where the last line follows from the definition of $\overline{\nabla}^{(t)}$ (cf. (58)). To continue, by the update rule in (57), we have

$$\begin{aligned} \log\bar{\nu}^{(t+1)} &= (1-\eta_t\tau)\log\nu^{(t)} - \eta_t A^\top\bar{\mu}^{(t)} + c_1\cdot\mathbf{1} \\ &= (1-\eta_t\tau)\Big[(1-\eta_{t-1}\tau)\log\nu^{(t-1)} - \eta_{t-1}A^\top\bar{\mu}^{(t)} + c_2\cdot\mathbf{1}\Big] - \eta_t A^\top\bar{\mu}^{(t)} + c_1\cdot\mathbf{1} \\ &= (1-\eta_t\tau)\Big[\log\bar{\nu}^{(t)} + \eta_{t-1}A^\top\bar{\mu}^{(t-1)} - \eta_{t-1}A^\top\bar{\mu}^{(t)}\Big] - \eta_t A^\top\bar{\mu}^{(t)} + c^{(t)}\cdot\mathbf{1} \\ &= \log\bar{\nu}^{(t)} - \eta_t\overline{\nabla}^{(t)} + (1-\eta_t\tau)\eta_{t-1}A^\top(\bar{\mu}^{(t-1)} - \bar{\mu}^{(t)}) + c^{(t)}\cdot\mathbf{1}, \end{aligned} \tag{60}$$

where the first three steps result from the update rule of $\bar{\nu}^{(t+1)}$, $\nu^{(t)}$ and $\bar{\nu}^{(t)}$, respectively, and the last line follows from (58). Here, c_1, c_2, and $c^{(t)}$ are some normalization constants.[4] Rearranging terms allows us to rewrite $\overline{\nabla}^{(t)}$ as

$$\overline{\nabla}^{(t)} = \frac{1}{\eta_t}\Big(\log\bar{\nu}^{(t)} - \log\bar{\nu}^{(t+1)} + c^{(t)}\cdot\mathbf{1}\Big) + \frac{\eta_{t-1}}{\eta_t}(1-\eta_t\tau)A^\top(\bar{\mu}^{(t-1)} - \bar{\mu}^{(t)}). \tag{61}$$

Plugging (61) into (59), we have

$$\begin{aligned}
&f_\tau^{(t)}(\bar{\nu}^{(t)}) - f_\tau^{(t)}(\nu)\\
&= \frac{1}{\eta_t}\langle\log\bar{\nu}^{(t)} - \log\bar{\nu}^{(t+1)}, \bar{\nu}^{(t)} - \nu\rangle - \tau\mathsf{KL}\big(\nu\,\|\,\bar{\nu}^{(t)}\big) + \frac{\eta_{t-1}}{\eta_t}(1-\eta_t\tau)\langle A^\top(\bar{\mu}^{(t-1)} - \bar{\mu}^{(t)}), \bar{\nu}^{(t)} - \nu\rangle\\
&= \frac{1}{\eta_t}\Big[\mathsf{KL}\big(\nu\,\|\,\bar{\nu}^{(t)}\big) - \mathsf{KL}\big(\nu\,\|\,\bar{\nu}^{(t+1)}\big) + \mathsf{KL}\big(\bar{\nu}^{(t)}\,\|\,\bar{\nu}^{(t+1)}\big)\Big] - \tau\mathsf{KL}\big(\nu\,\|\,\bar{\nu}^{(t)}\big)\\
&\quad + \frac{\eta_{t-1}}{\eta_t}(1-\eta_t\tau)\langle A^\top(\bar{\mu}^{(t-1)} - \bar{\mu}^{(t)}), \bar{\nu}^{(t)} - \nu\rangle.
\end{aligned}$$

Summing the equality over $t = 0, \ldots, T$ gives

$$\begin{aligned}
&\sum_{t=0}^{T}\Big[f_\tau^{(t)}(\bar{\nu}^{(t)}) - f_\tau^{(t)}(\nu)\Big]\\
&= \left(\frac{1}{\eta_0} - \tau\right)\mathsf{KL}\big(\nu\,\|\,\bar{\nu}^{(0)}\big) + \sum_{t=1}^{T}\left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} - \tau\right)\mathsf{KL}\big(\nu\,\|\,\bar{\nu}^{(t)}\big)\\
&\quad + \sum_{t=0}^{T}\frac{1}{\eta_t}\mathsf{KL}\big(\bar{\nu}^{(t)}\,\|\,\bar{\nu}^{(t+1)}\big) + \sum_{t=1}^{T}\frac{\eta_{t-1}}{\eta_t}(1-\eta_t\tau)\langle A^\top(\bar{\mu}^{(t-1)} - \bar{\mu}^{(t)}), \bar{\nu}^{(t)} - \nu\rangle.
\end{aligned} \tag{62}$$

With the choice of the learning rate

$$\eta_t = \frac{1}{(t+1)\tau},$$

one has

$$\frac{1}{\eta_0} - \tau = 0, \qquad \frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} - \tau = 0, \qquad \frac{\eta_{t-1}}{\eta_t}(1-\eta_t\tau) = 1,\ \forall\, t \geq 1. \tag{63}$$

Plugging the above relations into (62) leads to

$$\sum_{t=0}^{T}\Big[f_\tau^{(t)}(\bar{\nu}^{(t)}) - f_\tau^{(t)}(\nu)\Big] = \sum_{t=0}^{T}\frac{1}{\eta_t}\mathsf{KL}\big(\bar{\nu}^{(t)}\,\|\,\bar{\nu}^{(t+1)}\big) + \sum_{t=1}^{T}\langle A^\top(\bar{\mu}^{(t-1)} - \bar{\mu}^{(t)}), \bar{\nu}^{(t)} - \nu\rangle. \tag{64}$$

Next, we seek to bound the two terms in (64) separately.

4. We shall set $\eta_{-1} = 0$ and $\bar{\mu}^{(-1)} = \bar{\mu}^{(0)}$ to accommodate the case when $t = 0$ in (60).

Bounding the first term of the regret. According to Lemma 13 and (60), we have

$$\begin{aligned}
\mathsf{KL}\big(\bar{\nu}^{(t)} \,\|\, \bar{\nu}^{(t+1)}\big) &\le \frac{1}{2}\left\|\log \bar{\nu}^{(t+1)} - \log \bar{\nu}^{(t)} - c^{(t)}\cdot \mathbf{1}\right\|_{\infty}^2 \\
&= \frac{1}{2}\left\|-\eta_t \overline{\nabla}^{(t)} + (1-\eta_t\tau)\eta_{t-1}A^{\top}(\bar{\mu}^{(t-1)} - \bar{\mu}^{(t)})\right\|_{\infty}^2 \\
&\le \left\|-\eta_t \overline{\nabla}^{(t)}\right\|_{\infty}^2 + \left\|(1-\eta_t\tau)\eta_{t-1}A^{\top}(\bar{\mu}^{(t-1)} - \bar{\mu}^{(t)})\right\|_{\infty}^2 \\
&\le {\eta_t}^2(\tau\log|\mathcal{B}| + 3\left\|A\right\|_{\infty})^2 + 4{\eta_t}^2\left\|A\right\|_{\infty}^2,
\end{aligned}$$

where the final step results from Lemma 15 and the last equality in (63). Summing the above inequality over $t = 0, \cdots, T$ yields

$$\begin{aligned}
\sum_{t=0}^{T}\frac{1}{\eta_t}\mathsf{KL}\big(\bar{\nu}^{(t)} \,\|\, \bar{\nu}^{(t+1)}\big) &\le \sum_{t=0}^{T}\eta_t(\tau\log|\mathcal{B}| + 3\left\|A\right\|_{\infty})^2 + 4\sum_{t=0}^{T}\eta_t\left\|A\right\|_{\infty}^2 \\
&\le \tau^{-1}(\log T + 1)\left[(\tau\log|\mathcal{B}| + 3\left\|A\right\|_{\infty})^2 + 4\left\|A\right\|_{\infty}^2\right], \qquad (65)
\end{aligned}$$

where the last line follows from $\sum_{t=0}^{T}\eta_t \le \tau^{-1}(\log T + 1)$ due to the choice of the learning rate.

Bounding the second term of the regret. Observe that by the telescoping relation, we have

$$\begin{aligned}
&\sum_{t=1}^{T}\langle A^{\top}(\bar{\mu}^{(t-1)} - \bar{\mu}^{(t)}), \bar{\nu}^{(t)} - \nu\rangle \\
&= \sum_{t=1}^{T}\langle A^{\top}(\bar{\mu}^{(t-1)} - \bar{\mu}^{(t)}), \bar{\nu}^{(t)}\rangle - \langle A^{\top}(\mu^{(0)} - \bar{\mu}^{(T)}), \nu\rangle \\
&= \sum_{t=1}^{T}\langle A^{\top}\bar{\mu}^{(t-1)}, \bar{\nu}^{(t)} - \bar{\nu}^{(t-1)}\rangle + \sum_{t=1}^{T}\left[\langle A^{\top}\bar{\mu}^{(t-1)}, \bar{\nu}^{(t-1)}\rangle - \langle A^{\top}\bar{\mu}^{(t)}, \bar{\nu}^{(t)}\rangle\right] - \langle A^{\top}(\mu^{(0)} - \bar{\mu}^{(T)}), \nu\rangle \\
&= \sum_{t=1}^{T}\langle A^{\top}\bar{\mu}^{(t-1)}, \bar{\nu}^{(t)} - \bar{\nu}^{(t-1)}\rangle + \langle A^{\top}\mu^{(0)}, \bar{\nu}^{(0)}\rangle - \langle A^{\top}\bar{\mu}^{(T)}, \bar{\nu}^{(T)}\rangle - \langle A^{\top}(\mu^{(0)} - \bar{\mu}^{(T)}), \nu\rangle \\
&\le \left\|A\right\|_{\infty}\sum_{t=0}^{T-1}\left\|\bar{\nu}^{(t+1)} - \bar{\nu}^{(t)}\right\|_1 + 4\left\|A\right\|_{\infty}. \qquad (66)
\end{aligned}$$

Due to Pinsker's inequality and Lemma 13, we have

$$\frac{1}{2}\left\|\bar{\nu}^{(t+1)} - \bar{\nu}^{(t)}\right\|_1^2 \le \mathsf{KL}\big(\bar{\nu}^{(t+1)} \,\|\, \bar{\nu}^{(t)}\big) \le \frac{1}{2}\left\|\log \bar{\nu}^{(t+1)} - \log \bar{\nu}^{(t)} - c^{(t)}\cdot \mathbf{1}\right\|_{\infty}^2,$$

which further ensures that

$$\begin{aligned}
\left\|\bar{\nu}^{(t+1)} - \bar{\nu}^{(t)}\right\|_1 &\le \left\|\log \bar{\nu}^{(t+1)} - \log \bar{\nu}^{(t)} - c^{(t)}\cdot \mathbf{1}\right\|_{\infty} \\
&= \left\|-\eta_t \overline{\nabla}^{(t)} + (1-\eta_t\tau)\eta_{t-1}A^{\top}(\bar{\mu}^{(t-1)} - \bar{\mu}^{(t)})\right\|_{\infty}
\end{aligned}$$

$$\leq \eta_t \left\|\overline{\nabla}^{(t)}\right\|_\infty + \eta_t \left\|A^\top(\bar{\mu}^{(t-1)} - \bar{\mu}^{(t)})\right\|_\infty$$
$$\leq \eta_t \left(\tau \log|\mathcal{B}| + 5\|A\|_\infty\right),$$

where the second line follows from (60), the third line follows from the triangle inequality, and the last line follows from Lemma 15. Plugging the above inequality into (66) leads to

$$\sum_{t=1}^{T} \langle A^\top(\bar{\mu}^{(t-1)} - \bar{\mu}^{(t)}), \bar{\nu}^{(t)} - \nu\rangle \leq \tau^{-1}(\log T + 1)\|A\|_\infty \left(\tau \log|\mathcal{B}| + 5\|A\|_\infty\right) + 4\|A\|_\infty, \tag{67}$$

where we use again $\sum_{t=0}^{T} \eta_t \leq \tau^{-1}(\log T + 1)$.

Putting things together. Combining (65) and (67) into (64), we have

$$\mathsf{Regret}_\lambda(T) = \max_{\nu\in\Delta(\mathcal{B})} \sum_{t=0}^{T} \left[f_\tau^{(t)}(\nu^{(t)}) - f_\tau^{(t)}(\nu)\right] \leq \tau^{-1}(\log T + 1)(\tau \log|\mathcal{B}| + 5\|A\|_\infty)^2 + 4\|A\|_\infty.$$

Appendix B. Analysis for entropy-regularized Markov games

B.1 Proof of Proposition 2

For each t, let

$$V^{(t)}(s) := \max_{\mu(s)\in\Delta(\mathcal{A})} \min_{\nu(s)\in\Delta(\mathcal{B})} f_\tau(Q^{(t)}(s); \mu(s), \nu(s)),$$

which is, in other words, the minimax value of the associated matrix game using a payoff matrix $Q^{(t)}(s)$. We start by making a simple observation that for $\mu(s) \in \Delta(\mathcal{A}), \nu(s) \in \Delta(\mathcal{B})$,

$$\left|f_{Q^{(t)}(s)}(\mu(s),\nu(s)) - f_{Q_\tau^\star(s)}(\mu(s),\nu(s))\right| = \left|\mu(s)^\top (Q^{(t)}(s) - Q_\tau^\star(s))\nu(s)\right|$$
$$\leq \left\|Q^{(t)}(s) - Q_\tau^\star(s)\right\|_\infty \leq \left\|Q^{(t)} - Q_\tau^\star\right\|_\infty.$$

As a direct consequence, we can control $V^{(t)}(s) - V_\tau^\star(s)$ by

$$\left|V^{(t)}(s) - V_\tau^\star(s)\right|$$
$$= \left|\max_{\mu(s)\in\Delta(\mathcal{A})} \min_{\nu(s)\in\Delta(\mathcal{B})} f_{Q^{(t)}(s)}(\mu(s),\nu(s)) - \max_{\mu(s)\in\Delta(\mathcal{A})} \min_{\nu(s)\in\Delta(\mathcal{B})} f_{Q_\tau^\star(s)}(\mu(s),\nu(s))\right|$$
$$\leq \left\|Q^{(t)} - Q_\tau^\star\right\|_\infty.$$

Recalling the definition of the soft Bellman operator $\mathcal{T}_\tau$ in (17), it then follows that

$$\left\|Q^{(t+1)} - Q_\tau^\star\right\|_\infty = \left\|\mathcal{T}_\tau(Q^{(t)}) - \mathcal{T}_\tau(Q_\tau^\star)\right\| = \gamma \left|\mathbb{E}_{s'\sim P(\cdot|s,a,b)}\left[V^{(t)}(s') - V_\tau^\star(s')\right]\right|$$
$$\leq \gamma \left\|V^{(t)} - V_\tau^\star\right\|_\infty \leq \gamma \left\|Q^{(t)} - Q_\tau^\star\right\|_\infty.$$

Recursively invoking the above inequality proves inequality (19).

B.2 Proof of Theorem 9

The inner loop of Algorithm 3 aims to solve an entropy-regularized matrix game indexed by $Q^{(t)}(s)$, which is done by running the proposed PU or OMWU methods. To analyze the efficacy of the inner loop, let us denote the exact minimax game value on state s at t-th iteration by

$$\breve{V}^{(t+1)}(s) := \max_{\mu(s)\in\Delta(\mathcal{A})} \min_{\nu(s)\in\Delta(\mathcal{A})} f_\tau(Q^{(t)}(s);\mu(s),\nu(s)), \tag{68}$$

which is adopted in the exact value iteration analyzed in Proposition 2, and achieved by the equilibrium $\zeta^{\star(t)} = (\mu^{\star(t)}, \nu^{\star(t)})$ of (68).

- Denote the output of the inner loop as $\bar{\zeta}^{(t,T_{\rm sub})} = (\bar{\mu}^{(t,T_{\rm sub})}(s), \bar{\nu}^{(t,T_{\rm sub})}(s))$, which the entropy-regularized matrix game (68) is approximately solved by executing PU / OMWU for $T_{\rm sub}$ iterations. Theorem 2 (cf. (9c)) guarantees that for every $s \in \mathcal{S}$, one has

$$\begin{aligned}\left|V^{(t+1)}(s) - \breve{V}^{(t+1)}(s)\right| &= \left|f^\tau_{Q^{(t)}(s)}(\bar{\mu}^{(t,T_{\rm sub})}(s), \bar{\nu}^{(t,T_{\rm sub})}(s)) - f^\tau_{Q^{(t)}(s)}(\mu^{\star(t)}(s), \nu^{\star(t)}(s))\right| \\ &\le \eta^{-1}\frac{1}{1-(\tau+\|Q^{(t)}(s)\|_\infty)\eta}\cdot\frac{(1-\eta\tau)^{T_{\rm sub}}}{1-(1-\eta\tau)^{T_{\rm sub}}}\cdot \mathsf{KL}\big(\zeta^{\star(t)}\,\|\,\zeta^{(0)}\big) \\ &\le 2\eta^{-1}\cdot\frac{(1-\eta\tau)^{T_{\rm sub}}}{1-(1-\eta\tau)^{T_{\rm sub}}}\cdot 2\log|\mathcal{A}|,\end{aligned}$$

  where the last step makes use of the choice of the learning rate

$$\eta = \frac{1-\gamma}{2(1+\tau(\log|\mathcal{A}|+1))} \le \frac{1}{2(\tau+\|Q^{(t)}(s)\|_\infty)}$$

  and $\mathsf{KL}\big(\zeta^{\star(t)}\,\|\,\zeta^{(0)}\big) \le \log|\mathcal{A}| + \log|\mathcal{B}| \le 2\log|\mathcal{A}|$. As a consequence, setting

$$T_{\rm sub} = O\left(\frac{1}{\eta\tau}\left(\log\frac{1}{\epsilon} + \log\frac{1}{1-\gamma} + \log\log|\mathcal{A}| + \log\frac{1}{\eta}\right)\right) \tag{69}$$

  yields

$$\left|V^{(t+1)}(s) - \breve{V}^{(t+1)}(s)\right| \le (1-\gamma)\epsilon, \quad \text{for all} \quad s\in\mathcal{S}. \tag{70}$$

- We now move to monitor the progress of the outer loop. Combining (70) with some basic calculations, we arrive at

$$\begin{aligned}\left\|Q^{(t+1)} - Q^\star_\tau\right\|_\infty \le \gamma\left\|V^{(t+1)} - V^\star_\tau\right\|_\infty &\le \gamma\left\|V^{(t+1)} - \breve{V}^{(t+1)}\right\|_\infty + \gamma\left\|\breve{V}^{(t+1)} - V^\star_\tau\right\|_\infty \\ &\le (1-\gamma)\epsilon + \gamma\left\|Q^{(t)} - Q^\star_\tau\right\|_\infty.\end{aligned}$$

  Now invoking the above relation recursively, it is ensured that

$$\left\|Q^{(t+1)} - Q^\star_\tau\right\|_\infty \le \epsilon + \gamma^{t+1}\left\|Q^{(0)} - Q^\star_\tau\right\|_\infty.$$

  In view of the above relation, if one takes

$$T_{\rm main} = O\left(\frac{1}{1-\gamma}\left(\log\frac{1}{\epsilon} + \log\frac{1+\tau\log|\mathcal{A}|}{1-\gamma}\right)\right) \tag{71}$$

  iterations of the outer loop in Algorithm 3, we have $\left\|Q^{(T_{\rm main})} - Q^\star_\tau\right\|_\infty \le 2\epsilon$ as desired.

Putting things together, the total iteration complexity sufficient to achieve ϵ-accuracy equals to

$$T_{\mathrm{main}}T_{\mathrm{sub}} = O\left(\frac{1}{\eta\tau(1-\gamma)}\left(\log\frac{1}{\epsilon}+\log\log|\mathcal{A}|+\log\frac{1}{1-\gamma}+\log\tau\right)\left(\log\frac{1}{\epsilon}+\log\log|\mathcal{A}|+\log\frac{1}{1-\gamma}+\log\frac{1}{\eta}\right)\right).$$

Therefore the advertised iteration complexity in Theorem 9 holds true by simply noticing that $\eta = \frac{1-\gamma}{2(1+\tau(\log(|\mathcal{A}|)+1))}$ and $\tau < 1$, and hence

$$\log\left(\frac{1}{\eta}\right) \le \log\left(\frac{2\log|\mathcal{A}|+4}{1-\gamma}\right), \qquad \text{and} \qquad \log(\tau) \le \log\left(\frac{1}{1-\gamma}\right).$$

B.3 Proof of Corollary 10

We begin by recording two supporting lemmas whose proofs are deferred to Appendix C.6 and Appendix C.7.

Lemma 16 *Let $\zeta_\tau^\star = (\mu_\tau^\star, \nu_\tau^\star)$ be the QRE of payoff matrix A and $\widetilde{\zeta}_\tau^\star = (\widetilde{\mu}_\tau^\star, \widetilde{\nu}_\tau^\star)$ be that of $\widetilde{A}$. We have*

$$\left\|\log\zeta_\tau^\star - \log\widetilde{\zeta}_\tau^\star\right\|_\infty \le \frac{2}{\tau}\cdot\left(\frac{2}{\tau}\|A\|_\infty + 1\right)\left\|A-\widetilde{A}\right\|_\infty.$$

Lemma 17 *For any single-agent MDP $(\mathcal{S},\mathcal{A},P,r,\gamma)$ with bounded reward $0\le r\le R$, the entropy-regularized value function satisfies*

$$\left|V_\tau^{\pi'}(\rho) - V_\tau^{\pi}(\rho)\right| \le \frac{1}{1-\gamma}\left[\frac{R}{1-\gamma}+\tau\left(\frac{\log|\mathcal{A}|}{1-\gamma}+1+\log(|\mathcal{A}|+1)\right)\right]\left\|\log\pi'-\log\pi\right\|_\infty$$

for any two policies π and π'.

We are now ready to prove Corollary 10, which we break into a few steps.

Step 1: iteration complexity to obtain an approximate $Q_\tau^\star$. It is immediate from Theorem 9 that $T_{\mathsf{Q}} = \widetilde{O}\big(\frac{1}{\tau(1-\gamma)^2}\log^2(\frac{1}{\epsilon_{\mathsf{Q}}})\big)$ iterations are sufficient to get a $\widetilde{Q}_\tau^\star \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|\times|\mathcal{B}|}$ that achieves

$$\left\|\widetilde{Q}_\tau^\star - Q_\tau^\star\right\|_\infty \le \epsilon_{\mathsf{Q}}.$$

Step 2: iteration complexity to obtain an approximate $\zeta_\tau^\star$. Denote the QRE of the matrix game induced by $\widetilde{Q}_\tau^\star$ by $\widetilde{\zeta}_\tau^\star = (\widetilde{\mu}_\tau^\star, \widetilde{\nu}_\tau^\star)$. Specifically, for every $s\in\mathcal{S}$, $(\widetilde{\mu}_\tau^\star(s), \widetilde{\nu}_\tau^\star(s))$ solves the entropy-regularized matrix game induced by $\widetilde{Q}_\tau^\star(s)$. Invoking PU or OMWU with $\eta = \frac{1-\gamma}{2(1+\tau(\log|\mathcal{A}|+1-\gamma))}$ ensures that within $T_{\mathsf{policy}} = \widetilde{O}\big(\frac{1+\tau\log|\mathcal{A}|}{(1-\gamma)\tau}\log\frac{1}{\epsilon_{\mathsf{policy}}}\big)$ iterations, we can find a policy pair $\zeta = (\mu,\nu)$ such that

$$\left\|\log\zeta - \log\widetilde{\zeta}_\tau^\star\right\|_\infty \le \epsilon_{\mathsf{policy}}.$$

This taken together with Lemma 16 gives

$$\|\log\zeta - \log\zeta_\tau^\star\|_\infty \le \left\|\log\zeta - \log\widetilde{\zeta}_\tau^\star\right\|_\infty + \left\|\log\zeta_\tau^\star - \log\widetilde{\zeta}_\tau^\star\right\|_\infty$$

$$\leq \epsilon_{\mathsf{policy}} + \frac{2}{\tau} \cdot \left(\frac{2}{\tau} \left\| Q_\tau^\star \right\|_\infty + 1 \right) \left\| Q_\tau^\star - \widetilde{Q}_\tau^\star \right\|_\infty$$
$$\leq \epsilon_{\mathsf{policy}} + \frac{4 + 4\tau \log |\mathcal{A}| + 2\tau}{\tau^2 (1-\gamma)} \cdot \epsilon_{\mathsf{Q}}.$$

Therefore, we can get $\left\| \log \zeta - \log \zeta_\tau^\star \right\|_\infty \leq \epsilon \cdot C^{-1}$ within

$$T_{\mathsf{Q}} + T_{\mathsf{policy}} = \widetilde{O}\left(\frac{1}{\tau(1-\gamma)^2} \log^2 (\frac{1}{\epsilon_{\mathsf{Q}}}) \right) + \widetilde{O}\left(\frac{1 + \tau \log |\mathcal{A}|}{\tau(1-\gamma)} \log \frac{1}{\epsilon_{\mathsf{policy}}} \right)$$
$$= \widetilde{O}\left(\frac{1}{\tau(1-\gamma)^2} \log^2 (\frac{1}{\epsilon}) \right)$$

iterations as long as $C = \mathsf{poly}\left((1-\gamma)^{-1}, \tau^{-1}, \log |\mathcal{A}| \right)$.

Step 3: bounding the dual gap. For any μ', ν' we have

$$V_\tau^{\mu',\nu}(\rho) - V_\tau^{\mu,\nu'}(\rho) = \left(V_\tau^{\mu',\nu}(\rho) - V_\tau^{\mu_\tau^\star,\nu_\tau^\star}(\rho) \right) + \left(V_\tau^{\mu_\tau^\star,\nu_\tau^\star}(\rho) - V_\tau^{\mu,\nu'}(\rho) \right)$$
$$\leq \left(V_\tau^{\mu',\nu}(\rho) - V_\tau^{\mu',\nu_\tau^\star}(\rho) \right) + \left(V_\tau^{\mu_\tau^\star,\nu'}(\rho) - V_\tau^{\mu,\nu'}(\rho) \right), \tag{72}$$

where in each term, only one of the policies is varied. Consequently, it is possible to invoke the well-known performance difference lemma for single-agent MDP.

We note that the same policy μ' appears in both $V_\tau^{\mu',\nu}(\rho)$ and $V_\tau^{\mu',\nu_\tau^\star}(\rho)$, and it is therefore possible to invoke performance difference lemma for single-agent MDP to characterize $V_\tau^{\mu',\nu}(\rho) - V_\tau^{\mu',\nu_\tau^\star}(\rho)$, we construct a MDP $(\mathcal{S}, \mathcal{B}, \overline{P}, \overline{r}, \gamma)$ with

$$\overline{P}(s'|s,b) = \sum_{a \in \mathcal{A}} \mu'(a|s) P(s'|s,a,b),$$
$$\overline{r}(s,b) = \sum_{a \in \mathcal{A}} \mu'(a|s)(r(s,a,b) - \tau \log \mu'(a|s)),$$

and denote the associated entropy-regularized value function by $\overline{V}_\tau$. This allows us to write $V_\tau^{\mu',\nu}(\rho) = \overline{V}_\tau^{\nu}(\rho)$ and $V_\tau^{\mu',\nu_\tau^\star}(\rho) = \overline{V}_\tau^{\nu_\tau^\star}(\rho)$ (cf. (15)). Applying Lemma 17 with $R = 1 + \tau \log |\mathcal{A}|$ gives

$$\left| V_\tau^{\mu',\nu}(\rho) - V_\tau^{\mu',\nu_\tau^\star}(\rho) \right|$$
$$= \left| \overline{V}_\tau^{\nu}(\rho) - \overline{V}_\tau^{\nu_\tau^\star}(\rho) \right|$$
$$\leq \frac{1}{1-\gamma} \left[\frac{1 + \tau \log |\mathcal{A}|}{1-\gamma} + \tau \left(\frac{\log |\mathcal{B}|}{1-\gamma} + 1 + \log(|\mathcal{B}| + 1) \right) \right] \left\| \log \nu - \log \nu_\tau^\star \right\|_\infty.$$

Similarly, one can derive

$$\left| V_\tau^{\mu,\nu'}(\rho) - V_\tau^{\mu_\tau^\star,\nu'}(\rho) \right|$$
$$\leq \frac{1}{1-\gamma} \left[\frac{1 + \tau \log |\mathcal{B}|}{1-\gamma} + \tau \left(\frac{\log |\mathcal{A}|}{1-\gamma} + 1 + \log(|\mathcal{A}| + 1) \right) \right] \left\| \log \mu - \log \mu_\tau^\star \right\|_\infty$$

for the second term in (72). Plugging the above two inequalities into (72) gives

$$V_\tau^{\mu',\nu}(\rho) - V_\tau^{\mu,\nu'}(\rho) \le C \left\|\log\zeta - \log\zeta_\tau^\star\right\|_\infty \le \epsilon,$$

where the last inequality holds as long as $\left\|\log\zeta - \log\zeta_\tau^\star\right\|_\infty \le \epsilon \cdot C^{-1}$ with

$$C = \frac{2}{1-\gamma}\left[\frac{1}{1-\gamma} + \tau\left(\frac{\log|\mathcal{A}| + \log|\mathcal{B}|}{1-\gamma} + 1 + \log(|\mathcal{A}|+1) + \log(|\mathcal{B}|+1)\right)\right].$$

Therefore it takes $\widetilde{O}\big(\frac{1}{\tau(1-\gamma)^2}\log^2(\frac{1}{\epsilon})\big)$ iterations to achieve

$$\max_{\mu'} V_\tau^{\mu',\nu}(\rho) - \min_{\nu'} V_\tau^{\mu,\nu'}(\rho) \le \epsilon.$$

Appendix C. Proof of auxiliary lemmas

C.1 Proof of Lemma 11

Lemma 11 follows directly from the update sequence (6) and the form of the optimal solution pair $(\mu_\tau^\star, \nu_\tau^\star)$, provided in (4). Given the update sequence (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

$$\left\langle \log\mu^{(t+1)} - (1-\eta\tau)\log\mu^{(t)}, z_1 - \mu_\tau^\star\right\rangle = \eta z_1^\top A z_2 - \eta\mu_\tau^{\star\top} A z_2, \tag{73}$$

Similarly, one can derive

$$\left\langle \log\nu^{(t+1)} - (1-\eta\tau)\log\nu^{(t)}, z_2 - \nu_\tau^\star\right\rangle = -\eta z_1^\top A z_2 + \eta z_1^\top A\nu_\tau^\star. \tag{74}$$

By summing up equations (73) and (74), it is guarantee that

$$\left\langle \log\zeta^{(t+1)} - (1-\eta\tau)\log\zeta^{(t)}, \zeta_z - \zeta_\tau^\star\right\rangle = -\eta\mu_\tau^{\star\top} A z_2 + \eta z_1^\top A\nu_\tau^\star, \tag{75}$$

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

$$\eta\tau\log\mu_\tau^\star = \eta A\nu_\tau^\star + c\cdot\mathbf{1}, \tag{76}$$

for some normalization constant c. Again, by taking the inner product with $z_1 - \mu_\tau^\star$, we have

$$\left\langle \eta\tau\log\mu_\tau^\star, z_1 - \mu_\tau^\star\right\rangle = \eta(z_1 - \mu_\tau^\star)^\top A\nu_\tau^\star, \tag{77}$$

and similarly

$$\left\langle \eta\tau \log \nu_\tau^\star, z_2 - \nu_\tau^\star \right\rangle = \eta {\mu_\tau^\star}^\top A(z_2 - \nu_\tau^\star). \tag{78}$$

Combining inequalities (73) and (77), we arrive at inequality (23a); combining inequalities (74) and (78) gives inequality (23b). Moreover, putting together inequalities (75), (77) and (78) 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 Proof of Lemma 12

We begin with establishing (25a). By the definition of $f_\tau(\mu, \nu)$, direct calculations yield

$$\begin{aligned} f_\tau(\mu_\tau^\star, \nu_\tau^\star) - f_\tau(\mu, \nu_\tau^\star) &= (\mu_\tau^\star - \mu)^\top A \nu_\tau^\star + \tau \mu^\top \log \mu - \tau {\mu_\tau^\star}^\top \log \mu_\tau^\star \\ &= \tau \left(\left\langle \mu_\tau^\star - \mu, \log \mu_\tau^\star \right\rangle + \mu^\top \log \mu - {\mu_\tau^\star}^\top \log \mu_\tau^\star \right) = \tau \mathsf{KL}\big(\mu \,\|\, \mu_\tau^\star\big). \end{aligned} \tag{79}$$

Here, the second equality is obtained by plugging in (76). Similarly, we have

$$f_\tau(\mu_\tau^\star, \nu) - f_\tau(\mu_\tau^\star, \nu_\tau^\star) = \tau \mathsf{KL}\big(\nu \,\|\, \nu_\tau^\star\big). \tag{80}$$

Summing these two equalities completes the proof of (25a).

Turning to (25b), 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

$$f_\tau(\mu, \nu) - f_\tau(\mu_\tau^\star, \nu_\tau^\star) = (\mu_\tau^\star - \mu)^\top A(\nu_\tau^\star - \nu) + f_\tau(\mu_\tau^\star, \nu) + f_\tau(\mu, \nu_\tau^\star) - 2 f_\tau(\mu_\tau^\star, \nu_\tau^\star). \tag{81}$$

Finally, plugging (79) and (80) into (81) reveals the desired relation (25b).

C.3 Proof of Lemma 13

The second inequality follows directly from (Mei et al., 2020, Lemma 27). The first inequality has appeared, e.g., in Cen et al. (2022b). We reproduce a short proof for self-completeness. By straightforward calculations, the gradient of the function $\log(\|\exp(x)\|_1)$ is given by

$$\nabla_x \log(\|\exp(x)\|_1) = \exp(x) / \|\exp(x)\|_1,$$

which implies $\|\nabla_x \log(\|\exp(x)\|_1)\|_1 = 1$, $\forall x \in \mathbb{R}^{|\mathcal{A}|}$. Therefore, we have

$$\begin{aligned} \|\log \mu_1 - \log \mu_2\|_\infty &= \|x_1 - x_2 - \log(\|\exp(x_1)\|_1) \cdot \mathbf{1} + \log(\|\exp(x_2)\|_1) \cdot \mathbf{1}\|_\infty \\ &\le \|x_1 - x_2\|_\infty + \Big| - \log(\|\exp(x_1)\|_1) + \log(\|\exp(x_2)\|_1) \Big| \end{aligned}$$

$$
\begin{aligned}
&= \|x_1 - x_2\|_\infty + \Big| \langle x_1 - x_2, \nabla_x \log(\|\exp(x)\|_1)|_{x=x_c} \rangle \Big| \\
&\le \|x_1 - x_2\|_\infty + \Big| \|x_1 - x_2\|_\infty \|\nabla_x \log(\|\exp(x)\|_1)|_{x=x_c}\|_1 \Big| \\
&= 2\|x_1 - x_2\|_\infty ,
\end{aligned}
$$

where x_c is a certain convex combination of x_1 and x_2.

C.4 Proof of Lemma 14

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

$$
\begin{aligned}
f_\tau(\mu', \nu) - f_\tau(\mu, \nu') &= \big(f_\tau(\mu', \nu) - f_\tau(\mu', \nu_\tau^\star) - f_\tau(\mu, \nu') + f_\tau(\mu_\tau^\star, \nu')\big) - \big(f_\tau(\mu_\tau^\star, \nu') - f_\tau(\mu', \nu_\tau^\star)\big) \\
&= \big(f_\tau(\mu', \nu) - f_\tau(\mu', \nu_\tau^\star) - f_\tau(\mu, \nu') + f_\tau(\mu_\tau^\star, \nu')\big) - \tau \mathsf{KL}\big(\zeta' \,\|\, \zeta_\tau^\star\big), \quad (82)
\end{aligned}
$$

where the last step is due to $f_\tau(\mu, \nu_\tau^\star) - f_\tau(\mu_\tau^\star, \nu) = \tau \mathsf{KL}\big(\zeta \,\|\, \zeta_\tau^\star\big)$, as revealed in Lemma 12 (cf. (25a)).

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

$$
-f_\tau(\mu, \nu') + f_\tau(\mu_\tau^\star, \nu') = -(\mu - \mu_\tau^\star)^\top A(\nu' - \nu_\tau^\star) + f_\tau(\mu_\tau^\star, \nu_\tau^\star) - f_\tau(\mu, \nu_\tau^\star).
$$

Plugging the above two equalities into (82) gives

$$
\begin{aligned}
&f_\tau(\mu', \nu) - f_\tau(\mu, \nu') \\
&= \big(\mu' - \mu_\tau^\star\big)^\top A(\nu - \nu_\tau^\star) - (\mu - \mu_\tau^\star)^\top A(\nu' - \nu_\tau^\star) + f_\tau(\mu_\tau^\star, \nu) - f_\tau(\mu, \nu_\tau^\star) - \tau \mathsf{KL}\big(\zeta' \,\|\, \zeta_\tau^\star\big) \\
&= \big(\mu' - \mu_\tau^\star\big)^\top A(\nu - \nu_\tau^\star) - (\mu - \mu_\tau^\star)^\top A(\nu' - \nu_\tau^\star) + \tau \mathsf{KL}\big(\zeta \,\|\, \zeta_\tau^\star\big) - \tau \mathsf{KL}\big(\zeta' \,\|\, \zeta_\tau^\star\big) \\
&\le \|A\|_\infty \left(\big\|\mu' - \mu_\tau^\star\big\|_1 \|\nu - \nu_\tau^\star\|_1 + \big\|\nu' - \nu_\tau^\star\big\|_1 \|\mu - \mu_\tau^\star\|_1\right) + \tau \mathsf{KL}\big(\zeta \,\|\, \zeta_\tau^\star\big) - \tau \mathsf{KL}\big(\zeta' \,\|\, \zeta_\tau^\star\big) \\
&\overset{\text{(i)}}{\le} \frac{1}{2}\|A\|_\infty \left[\frac{\tau}{\|A\|_\infty}\left(\big\|\mu' - \mu_\tau^\star\big\|_1^2 + \big\|\nu' - \nu_\tau^\star\big\|_1^2\right) + \frac{\|A\|_\infty}{\tau}\left(\|\mu - \mu_\tau^\star\|_1^2 + \|\nu - \nu_\tau^\star\|_1^2\right)\right] \\
&\qquad + \tau \mathsf{KL}\big(\zeta \,\|\, \zeta_\tau^\star\big) - \tau \mathsf{KL}\big(\zeta' \,\|\, \zeta_\tau^\star\big) \\
&\overset{\text{(ii)}}{\le} \tau \mathsf{KL}\big(\zeta' \,\|\, \zeta_\tau^\star\big) + \frac{\|A\|_\infty^2}{\tau} \mathsf{KL}\big(\zeta_\tau^\star \,\|\, \zeta\big) + \tau \mathsf{KL}\big(\zeta \,\|\, \zeta_\tau^\star\big) - \tau \mathsf{KL}\big(\zeta' \,\|\, \zeta_\tau^\star\big) \\
&= \frac{\|A\|_\infty^2}{\tau} \mathsf{KL}\big(\zeta_\tau^\star \,\|\, \zeta\big) + \tau \mathsf{KL}\big(\zeta \,\|\, \zeta_\tau^\star\big),
\end{aligned}
$$

where the second step invokes Lemma 12 (cf. (25a)), (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.

C.5 Proof of Lemma 15

First, we show that the update of $\bar{\nu}^{(t)}$ (cf. (57b)) satisfies

$$\bar{\nu}^{(t)}(b) \propto \exp([A\widetilde{\mu}^{(t)}]_b/\tau) \quad \text{for some } \widetilde{\mu}^{(t)} \in \Delta(\mathcal{A}) \tag{83}$$

by induction.

- For $t = 0$, it is easily seen that $\bar{\nu}^{(0)}(b) = \frac{1}{|\mathcal{B}|} \propto \exp([A^\top \widetilde{\mu}^{(0)}]_b/\tau)$ with $\widetilde{\mu}^{(0)} = 0$.
- Now assume (83) holds for all steps up to t. The update rule (cf. (57b)) implies

$$\begin{aligned}\bar{\nu}^{(t+1)}(b) &\propto \bar{\nu}^{(t)}(b)^{1-\eta_t\tau} \exp(\eta_t [A^\top \bar{\mu}^{(t)}]) \\ &\propto \exp((1-\eta_t\tau)[A^\top \widetilde{\mu}^{(t)}]_b/\tau) \exp(\eta_t [A^\top \bar{\mu}^{(t)}]) = \exp([A^\top \widetilde{\mu}^{(t+1)}]_b/\tau),\end{aligned}$$

  with $\widetilde{\mu}^{(t+1)} = (1-\eta_t\tau)\widetilde{\mu}^{(t)} + \eta_t\tau\bar{\mu}^{(t)} \in \Delta(\mathcal{B})$.

Therefore, the claim (83) holds for all $t > 0$.

It then follows from (83) straightforwardly that

$$\frac{\bar{\nu}^{(t)}(b_1)}{\bar{\nu}^{(t)}(b_2)} = \frac{\exp([A\widetilde{\mu}^{(t)}]_{b_1}/\tau)}{\exp([A\widetilde{\mu}^{(t)}]_{b_2}/\tau)} \le \exp(2\left\|A\right\|_\infty/\tau)$$

for any $b_1, b_2 \in \mathcal{B}$. Therefore, we have

$$1 = \sum_{b\in\mathcal{B}} \bar{\nu}^{(t)}(b) \le |\mathcal{B}| \exp(2\left\|A\right\|_\infty/\tau) \cdot \min_{b\in\mathcal{B}} \bar{\nu}^{(t)}(b),$$

which gives $\min_{b\in\mathcal{B}} \bar{\nu}^{(t)}(b) \ge |\mathcal{B}|^{-1} \exp(-2\left\|A\right\|_\infty/\tau)$, or equivalently

$$\left\|\log \bar{\nu}^{(t)}\right\|_\infty \le 2\left\|A\right\|_\infty/\tau + \log|\mathcal{B}|.$$

We conclude the proof in view of the expression of $\overline{\nabla}^{(t)}$ in (58):

$$\left\|\overline{\nabla}^{(t)}\right\|_\infty = \left\|A^\top \bar{\mu}^{(t)} + \tau \log \bar{\nu}^{(t)}\right\|_\infty \le \left\|A^\top \bar{\mu}^{(t)}\right\|_\infty + \tau \left\|\log \nu^{(t)}\right\|_\infty \le \tau \log|\mathcal{B}| + 3\left\|A\right\|_\infty.$$

C.6 Proof of Lemma 16

Instantiating (77) at $z_1 := \mu_\tau'^\star$, we have

$$\left\langle \tau \log \mu_\tau^\star, \mu_\tau'^\star - \mu_\tau^\star \right\rangle = {\mu_\tau'^\star}^\top A \nu_\tau^\star - {\mu_\tau^\star}^\top A \nu_\tau^\star.$$

Similarly, instantiating (78) with $z_2 := \nu_\tau'^\star$, we obtain

$$\left\langle \tau \log \nu_\tau^\star, \nu_\tau'^\star - \nu_\tau^\star \right\rangle = -{\mu_\tau^\star}^\top A \nu_\tau'^\star + {\mu_\tau^\star}^\top A \nu_\tau^\star.$$

Summing the above two equalities then leads to

$$\tau\left\langle \log \zeta_\tau^\star, \zeta_\tau'^\star - \zeta_\tau^\star \right\rangle = -{\mu_\tau^\star}^\top A \nu_\tau'^\star + {\mu_\tau'^\star}^\top A \nu_\tau^\star.$$

In view of symmetry, the following equality holds as well

$$\tau\big\langle \log \zeta_\tau'^\star, \zeta_\tau^\star - \zeta_\tau'^\star\big\rangle = -\mu_\tau'^{\star\top} A' \nu_\tau^\star + \mu_\tau^{\star\top} A' \nu_\tau'^\star,$$

Taken the above relations collectively allows us to arrive at

$$\begin{aligned}
\mathsf{KL}\big(\zeta_\tau^\star \,\|\, \zeta_\tau'^\star\big) + \mathsf{KL}\big(\zeta_\tau'^\star \,\|\, \zeta_\tau^\star\big) &= \big\langle \log \zeta_\tau'^\star - \log \zeta_\tau^\star, \zeta_\tau'^\star - \zeta_\tau^\star\big\rangle \\
&= \frac{1}{\tau}\Big(\mu_\tau^{\star\top}(A-A')\nu_\tau'^\star - \mu'^{\star\top}(A-A')\nu_\tau^\star\Big) \\
&= \frac{1}{\tau}\Big(\mu_\tau^{\star\top}(A-A')(\nu_\tau'^\star - \nu_\tau^\star) - \big(\mu_\tau'^\star - \mu_\tau^\star\big)^\top (A-A')\nu_\tau^\star\Big) \\
&\le \frac{1}{\tau}\big\|A-A'\big\|_\infty \big(\big\|\mu_\tau^\star - \mu_\tau'^\star\big\|_1 + \big\|\nu_\tau^\star - \nu_\tau'^\star\big\|_1\big). \qquad (84)
\end{aligned}$$

On the other hand, we have

$$\frac{1}{2}\big(\big\|\mu_\tau^\star - \mu_\tau'^\star\big\|_1 + \big\|\nu_\tau^\star - \nu_\tau'^\star\big\|_1\big)^2 \le \big\|\mu_\tau^\star - \mu_\tau'^\star\big\|_1^2 + \big\|\nu_\tau^\star - \nu_\tau'^\star\big\|_1^2 \le \mathsf{KL}\big(\zeta_\tau^\star \,\|\, \zeta_\tau'^\star\big) + \mathsf{KL}\big(\zeta_\tau'^\star \,\|\, \zeta_\tau^\star\big), \qquad (85)$$

where the second step results from Pinsker's inequality. Combining (84) and (85) leads to

$$\big\|\zeta_\tau^\star - \zeta_\tau'^\star\big\|_1 = \big\|\mu_\tau^\star - \mu_\tau'^\star\big\|_1 + \big\|\nu_\tau^\star - \nu_\tau'^\star\big\|_1 \le \frac{2}{\tau}\big\|A-A'\big\|_\infty. \qquad (86)$$

This in turn allows us to bound

$$\begin{aligned}
\big\|A\nu_\tau^\star - A'\nu'^\star\big\|_\infty &\le \big\|A(\nu_\tau^\star - \nu'^\star)\big\|_\infty + \big\|(A-A')\nu'^\star\big\|_\infty \\
&\le \|A\|_\infty \big\|\nu_\tau^\star - \nu'^\star\big\|_1 + \big\|A-A'\big\|_\infty \\
&\le \|A\|_\infty \big\|\zeta_\tau^\star - \zeta_\tau'^\star\big\|_1 + \big\|A-A'\big\|_\infty \\
&\le \left(\frac{2}{\tau}\|A\|_\infty + 1\right)\big\|A-A'\big\|_\infty,
\end{aligned}$$

where the final step invokes (86). Since $\mu_\tau^\star \propto \exp(-A\nu_\tau^\star/\tau)$, $\mu'^\star \propto \exp(-A\nu'^\star/\tau)$, we invoke Lemma 13 to arrive at

$$\big\|\log \mu_\tau^\star - \log \mu'^\star\big\|_\infty \le \frac{2}{\tau}\big\|A\nu_\tau^\star - A'\nu'^\star\big\|_\infty \le \frac{2}{\tau}\cdot\left(\frac{2}{\tau}\|A\|_\infty + 1\right)\big\|A-A'\big\|_\infty,$$

which establishes the desired bound.

C.7 Proof of Lemma 17

Using the regularized version of the performance difference lemma (see (Zhan et al., 2021, Lemma 7) or (Lan, 2022, Lemma 2)), we have:

$$V_\tau^{\pi'}(\rho) - V_\tau^\pi(\rho) = \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_\rho^{\pi'}}\Big[\big\langle Q_\tau^{\pi'}(s), \pi'(\cdot|s) - \pi(\cdot|s)\big\rangle - \tau\big[\mathcal{H}(\pi'(\cdot|s)) - \mathcal{H}(\pi(\cdot|s))\big]\Big]. \qquad (87)$$

We then bound the two terms separately.

- To control the first term, we notice that $Q_\tau(s,a)$ is bounded by $0 \le Q_\tau(s,a) \le \frac{R}{1-\gamma} + \frac{\tau \log|\mathcal{A}|}{1-\gamma}$ for all (s,a). Hence, we have

$$\begin{aligned}\left|\left\langle Q_\tau^{\pi'}(s), \pi'(\cdot|s) - \pi(\cdot|s)\right\rangle\right| &\le \left\|Q_\tau^{\pi'}(s)\right\|_\infty \left\|\pi'(\cdot|s) - \pi(\cdot|s)\right\|_1 \\ &\le \left\|Q_\tau^{\pi'}(s)\right\|_\infty \left\|\log \pi'(\cdot|s) - \log \pi(\cdot|s)\right\|_\infty \\ &\le \frac{1}{1-\gamma}\left(R + \tau \log|\mathcal{A}|\right)\left\|\log \pi'(\cdot|s) - \log \pi(\cdot|s)\right\|_\infty, \quad (88)\end{aligned}$$

  where the second step is due to (Mei et al., 2020, Lemma 24).

- Turning to the entropy difference term in (87), let

$$F(\log \pi(\cdot|s)) = -\left\langle \exp(\log \pi(\cdot|s)), \log \pi(\cdot|s)\right\rangle = \mathcal{H}(\pi(\cdot|s)).$$

  We can then invoke mean value theorem to show

$$\begin{aligned}\left|-\mathcal{H}(\pi'(\cdot|s)) + \mathcal{H}(\pi(\cdot|s))\right| &= \left|-F(\log \pi'(\cdot|s)) + F(\log \pi(\cdot|s))\right| \\ &= \left|\left\langle \log \pi'(\cdot|s) - \log \pi(\cdot|s), \nabla F(\log \xi)\right\rangle\right| \\ &\le \left\|\log \pi'(\cdot|s) - \log \pi(\cdot|s)\right\|_\infty \left\|\nabla F(\log \xi)\right\|_1 \\ &\le \left\|\log \pi'(\cdot|s) - \log \pi(\cdot|s)\right\|_\infty \left(\|\xi\|_1 + \|\xi \odot \log \xi\|_1\right), \quad (89)\end{aligned}$$

  where $\log \xi = c \log \pi(\cdot|s) + (1-c)\log \pi'(\cdot|s)$ for some constant $0 < c < 1$, and the last line follows from $\nabla F(\log \xi) = -\xi - \xi \odot \log \xi$, where $\odot$ denotes point-wise multiplication. Hölder's inequality guarantees that

$$\|\xi\|_1 = \left|\left\langle \pi(\cdot|s)^c, \pi'(\cdot|s)^{1-c}\right\rangle\right| \le \|\pi(\cdot|s)^c\|_{1/c} \left\|\pi'(\cdot|s)^{1-c}\right\|_{1/(1-c)} = 1.$$

  Introduce $\overline{\xi}$ which appends a scalar $1 - \|\xi\|_1$ to ξ as $\overline{\xi} = [\xi^\top, 1 - \|\xi\|_1]^\top$, so that $\overline{\xi}$ is a probability vector. It is straightforward to get

$$\begin{aligned}\|\xi \odot \log \xi\|_1 &= -\sum_{a \in \mathcal{A}} \xi(a) \log \xi(a) \\ &\le -\sum_{a \in \mathcal{A}} \xi(a) \log \xi(a) - (1 - \|\xi\|_1)\log(1 - \|\xi\|_1) = \mathcal{H}(\overline{\xi}) \le \log(|\mathcal{A}| + 1).\end{aligned}$$

  Substitution of the above two inequalities into (89) gives

$$\left|-\mathcal{H}(\pi'(\cdot|s)) + \mathcal{H}(\pi(\cdot|s))\right| \le \left\|\log \pi'(\cdot|s) - \log \pi(\cdot|s)\right\|_\infty \left(1 + \log(|\mathcal{A}| + 1)\right). \quad (90)$$

Plugging (90) and (88) into (87) completes the proof.