On Solving Larger Games: Designing New Algorithms Adaptable to Deep Reinforcement Learning

by

Mingyang Liu

B.Eng., Tsinghua University (2023)

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2025

Authored by: Mingyang Liu
Department of Electrical Engineering and Computer Science
January 18, 2025

Certified by: Asuman Ozdaglar
MathWorks Professor of Electrical Engineering and Computer Science
Thesis Supervisor

Certified by: Gabriele Farina
Assistant Professor of Electrical Engineering and Computer Science
Thesis Supervisor

Accepted by: Leslie A. Kolodziejski
Professor of Electrical Engineering and Computer Science
Chair, Department Committee on Graduate Students

On Solving Larger Games: Designing New Algorithms Adaptable to Deep Reinforcement Learning

by

Mingyang Liu

Submitted to the Department of Electrical Engineering and Computer Science
on January 18, 2025 in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

ABSTRACT

In this thesis, we explore the design of algorithms capable of handling large games where the state space is too large to store strategies in a tabular format from a theoretical perspective. Specifically, we focus on developing algorithms suitable for deep reinforcement learning in two-player zero-sum extensive-form games. There are three critical properties for effective deep multi-agent reinforcement learning: (last/best) iterate convergence, efficient utilization of stochastic trajectory feedback, and theoretically sound avoidance of importance sampling corrections. Chapter 3 introduces Regularized Optimistic Mirror Descent (`Reg-OMD`), which provably converges to the Nash equilibrium (NE) linearly in last-iterate. Chapter 4 shows that algorithms based on regret decomposition enjoy best-iterate convergence to the NE. Chapter 5 proposes Q-value based Regret Minimization (`QFR`), which achieves all three properties simultaneously.

Thesis supervisor: Asuman Ozdaglar
Title: MathWorks Professor of Electrical Engineering and Computer Science

Thesis supervisor: Gabriele Farina
Title: Assistant Professor of Electrical Engineering and Computer Science

Acknowledgments

I wish to express my deep gratitude for the support and guidance provided by my advisors, Professor Asuman Ozdaglar and Professor Gabriele Farina. They gave me the freedom to explore research topics of interest and offered detailed guidance on research methodologies. Furthermore, I am continually motivated by Asuman's diligence throughout the enduring road of research and have learned a great deal from Gabriele's precise and meticulous approach to writing papers.

My gratitude extends to Kaiqing Zhang, who introduced me to the field of theoretical research and helped me establish a methodology for conducting it. Engaging with my labmates—Omar Bennouna, Steven DiSilvio, Fathima Zarin Faizal, Alireza Fallah, Zhiyuan Fan, Kimia Hassibi, Peter Hoffman, Kihyun Kim, Dingwen Kong, Charles Lyu, Gagik Magakyan, Sobhan Mohammadpour, Chanwoo Park, Sarath Pattathil, Charis Pipis, Ashkan Soleymani, and Evan Vogelbaum—has been incredibly rewarding. Discussions with them sparked innovation and consistently brought forth novel research ideas.

My academic journey began at Tsinghua University, where Professor Chongjie Zhang's mentorship and Noam Brown's talk introduced me to the world of research. This experience laid the foundation for my scholarly pursuits in game theory.

I am also grateful to the community provided by LIDS and EECS at MIT, which enabled me to exchange ideas and investigate intriguing topics. I look forward to contributing further to this community in the future.

Finally, the constant support and love from my parents have been my anchor throughout this journey. Their belief in my potential has continuously encouraged me to move forward.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Recently, deep reinforcement learning (DRL) has achieved great success in many fields, including games [Berner et al., 2019, Mnih et al., 2013, Silver et al., 2017, Vinyals et al., 2019], robotics [Ibarz et al., 2021], autonomous driving [Kiran et al., 2021], and large language models [Ouyang et al., 2022]. Most of these successes are based on scalable DRL algorithms, such as proximal policy optimization (PPO) [Schulman et al., 2017] and soft actor-critic (SAC) [Haarnoja et al., 2018], where scalability results from three key properties: (**I**) only requires value estimates obtained from repeated random rollouts, which can be implemented efficiently; (**II**) converges in iterates (as opposed to in averages), removing the need for either training an average policy approximator or storing snapshots of past policies; (**III**) soundly avoids importance sampling while estimating the reward of actions during sampling, which can be detrimental in practice as the resulting outsized reward estimates often lead to numerical instability.

Nonetheless, DRL cannot be directly applied to multi-agent imperfect information games, since its strategy will circle around the equilibrium without convergence [Balduzzi et al., 2019]. Furthermore, current theoretically sound algorithms in multi-agent imperfect information games are not suitable for DRL because of two obstacles: average-iterate convergence and importance sampling. We will discuss them in the following.

Basic Concepts and Terminologies

In this section, we will use the Texas hold'em as an example to illustrate the basic concepts and terminologies in games. There are multiple players in Texas hold'em, and each player will be dealt two private cards at the beginning of the game, and five public cards will be revealed to all players during the game. Note that only the owner has access to his private cards so his private cards are hidden from other players. Then, players will bet in turns until the game ends or all players except one quit the betting. For all the players that do not quit the game, the one with the largest card combination, composed of arbitrary five cards among his two private cards and five public cards, wins the game. In this paper, when Texas hold'em is mentioned, we refer to its two-player version, which is a two-player zero-sum game.

In Texas hold'em, the current state of the game, which constitutes the private cards of all

This chapter builds on Section 1 and Section 2 of Liu et al. [2023] and Liu et al. [2024b].

players, the public cards, and all betting sequences, is called the *game state*. It is also known as the *node* in the game tree since the whole game can be illustrated in a tree structure. Given the game tree of Texas hold'em, the root node represents the beginning of the game, where no private cards have been dealt and no players have bet. At any game state, the *Q-value* of taking an action (*i.e.,* betting in Texas hold'em) is the expected utility conditioned on taking this action.

Average-Iterate vs. Iterate Convergence

Most scalable previous works about solving Nash equilibrium (NE) in two-player zero-sum imperfect information extensive-form games (EFGs) are based on counterfactual regret minimization (CFR) [Zinkevich et al., 2007] and its variants [Brown and Sandholm, 2019a, Farina et al., 2019a, 2021a, Tammelin et al., 2015]. By applying no-regret learning on both players, the average strategy over all intermediate iterates generated by the no-regret learning algorithm will converge to the NE. There are two main ways to compute the average iterate in large games, where the strategies are approximated by neural networks.

Approximate Average Iterate by Neural Network. Since each iteration of the algorithm is already approximated by a neural network, which is inherently non-linear, averaging the network parameters to approximate their average strategy is inaccurate. Consequently, Brown et al. [2019] introduces an additional neural network for computing the average strategy. However, this approach introduces an additional layer of approximation error on top of the original error induced by the approximated strategy at each iteration. As a result, the algorithm's overall performance is adversely affected.

Sample Intermediate Iterates. An alternative way to approximate the average strategy is to store the intermediate iterates on disk, and sample one at test time [Steinberger, 2019, Steinberger et al., 2020]. Then, the expectation of the strategy is the average strategy while being sampled uniformly. Nonetheless, in a large game, it usually takes several gigabytes to store the parameters of a single network and millions of iterations to converge. Hence, it takes a large amount of memory. Furthermore, the variance of the average strategy's performance is larger due to the sampling.

Therefore, to resolve the issues above, recent works [Cen et al., 2021b, Daskalakis and Panageas, 2019, Lee et al., 2021, Liang and Stokes, 2019, Wei et al., 2021] focused on *iterate convergence*, *i.e.,* one of the intermediate iterate converges to the NE, to avoid averaging. There are two classes of iterate convergence: last-iterate convergence and best-iterate convergence. The main distinction is that last-iterate convergence ensures that the last iterate produced by the algorithm converges to the NE while best-iterate convergence only guarantees the existence of an intermediate iterate converges to the NE. Although the algorithms above satisfy iterate convergence, they are not compatible with DRL since they need either importance sampling or cannot sample efficiently.

Counterfactual Values v.s. Q-values

Due to the imperfect information, a player's strategy is a probability distribution over actions conditioned on *information sets* (or infosets for short), instead of the game states (a.k.a. nodes in the game tree). An infoset contains several game states, which only differ in the information hidden from the player at the infoset. For example, for player 1 in Texas hold'em, game states that differ from the private cards of player 2 are in the same infoset of player 1. Consequently, the expected utility of taking an action in an infoset is the *counterfactual value*, instead of the Q-value. The counterfactual value is essentially the Q-value multiplied by the probability of reaching the infoset contributed by the opponent's strategy.

Compared to Q-values, counterfactual values are asymmetric between players and thus require importance sampling to get unbiased estimators from rolling trajectories. We assume all randomness in the game resulted from the players' strategies for simplicity. For ease of presentation, we define $\Pr_i(s' \mid s, a)$ as an informal notation for the probability of reaching infoset s' conditioned on reaching the infoset-action pair (s, a), contributed by player $\{1, 2\} \ni i$'s strategy. In other words, the probability of reaching infoset s' from (s, a) is $\Pr_1(s' \mid s, a) \cdot \Pr_2(s' \mid s, a)$. Then, at infoset s, the counterfactual value of player 1 for taking action a is $\mathbb{E}_{s' \sim \Pr_1(\cdot|s,a) \Pr_2(\cdot|\emptyset)} [\mathcal{U}_1(s')]$, where $\emptyset$ represents the root of the game tree and $\mathcal{U}_1(s')$ is the utility of player 1 at infoset s'..

Nonetheless, both players' strategies contributed equally to the probability of a sampled trajectory starting from the root $\emptyset$ of the game. Therefore, to get an unbiased estimator for the counterfactual value of player 1 at infoset s, the reward collected from the trajectory need to be divided by $\Pr_1(s, a \mid \emptyset)$. Hence, the magnitude of the unbiased estimator for the counterfactual value is proportional to the reciprocal of player 1's reach probability to an infoset s, which can be as large as the game size [Bai et al., 2022, Fiegel et al., 2023, Kozuno et al., 2021, Lanctot et al., 2009]. For instance, this value can be as large as 10^{200} in Stratego, which will cause critical numerical instability of the neural network.

External Sampling v.s. Trajectory Rollouts

A potential remedy to the importance sampling is external sampling [Lanctot et al., 2009], which avoids importance sampling by enumerating all infosets and sampling a single node in each infoset. The main drawback of external sampling is that the number of infosets is still intractable in large games. In contrast, the number of nodes visited when sampling a trajectory is bounded by the height of the game tree, which is typically the logarithm of the number of nodes. For instance, external sampling needs to visit more than 10^{200} nodes in each iteration while rolling trajectory only visits no more than 4000 nodes in Stratego [Perolat et al., 2022].

1.1 Related Work

This section compares our thesis to previous works on four aspects: the use of regularization, convergence guarantees, the notion of values used by the algorithm, and support of stochastic feedback. We also illustrate the most relevant algorithms in Table 1.1.

Regularization

Overview. Additional regularization on the objective function has been widely used to accelerate convergence and encourage exploration in reinforcement learning, [Cen et al., 2021a, Geist et al., 2019, Mei et al., 2020, Tuyls et al., 2003]. In two-player zero-sum normal-form games (NFGs), the original bilinear objective can be turned to strongly convex-concave by applying additional regularization [Cen et al., 2021b, Hofbauer and Hopkins, 2005], where the game with additional regularization is also called the *regularized game*. Nonetheless, Hofbauer and Hopkins [2005] only gave asymptotic convergence to the NE of the regularized game under the best-response dynamics and Cen et al. [2021b] only provided convergence of optimistic multiplicative weight update (OMWU). Similar ideas could be dated back to the smoothing techniques led by Nesterov [2003]. There are also several papers that focus on continuous-time dynamics under additional regularization by utilizing Lyapunov arguments [Leonardos et al., 2021, Perolat et al., 2021].

By adding additional regularization, the convergence to the saddle point of the regularized game can be guaranteed. While the regularization is small, the solution to the regularized game can be close to the NE of the original problem, in terms of duality gap [Cen et al., 2021b].

Dilated Regularizer. To enable closed-form (efficient) updates of strategies, Hoda et al. [2010], Kroer et al. [2020] proposed the dilated regularizer, which weights the local regularizer at each infoset in the game tree by the player's reach probability. However, while computing an unbiased estimator with rolling trajectories, the additional dilated regularizer will result in importance sampling. Therefore, Liu et al. [2024b] proposed *bidilated regularizer* to avoid importance sampling.

Anchoring (Slingshot) Regularizer. Typically, prior works fixed the regularizer[1] throughout the game solving [Cen et al., 2021b, Liu et al., 2023, Mei et al., 2020]. For instance, the entropy of players' strategies might be added to the objective function throughout the process. Recently, some works [Abe et al., 2024, Bakhtin et al., 2022, Perolat et al., 2021, 2022] has focused on *anchoring regularizer*, which generalizes the original regularizer to its corresponding Bregman distance. For example, entropy on strategies is generalized to the KL divergence between the current strategy and a fixed strategy, which is called *anchoring strategy* [Abe et al., 2024]. When the anchoring strategy is chosen to be the uniform distribution over actions, the anchoring regularizer typically degenerates to the original regularizer.

Bakhtin et al. [2022] used the anchoring strategy to force the learned strategy close to the strategy imitating human game plays. Moreover, to get closer to the original game, shrinking the magnitude of the regularizer can be substituted by iteratively changing the anchoring strategy to the latest strategy produced by the algorithm. Since the iterations are converging, the anchoring regularizer will gradually converge to the distance to NE, and thus the NE of the regularized game will be close to the NE of the original game. The advantage of such method is that the learning algorithm is always solving a strongly convex-concave objective

[1] But they may shrink the magnitude of the regularization term through learning to get closer to the original objective.

with the convexity large enough (in contrast to shrink regularization magnitude), which is beneficial for accelerating and stabilizing deep reinforcement learning in practice [Perolat et al., 2022].

Convergence Guarantees.

Average-iterate convergence. Most variants of CFR [Steinberger et al., 2020, Tammelin et al., 2015, Zinkevich et al., 2007] only guarantee that the *average* of all intermediate strategies generated by the algorithm converges to an NE, though empirically some of them show last-iterate convergence [Bowling et al., 2015, Tammelin et al., 2015]. As stated in the introduction, the downside of average-iterate convergence is that it is not amenable to deep learning. Therefore, there is a recent trend focusing on last-iterate convergence.

Last-iterate convergence in NFGs. While the last iterate generated by mirror descent diverges [Bailey and Piliouras, 2018, Mertikopoulos et al., 2018], its optimistic version achieves great success in solving games, enabling both faster and last-iterate convergence guarantees [Cai et al., 2022, Daskalakis et al., 2018, Lei et al., 2021, Mertikopoulos et al., 2019, Mokhtari et al., 2020, Rakhlin and Sridharan, 2013]. While assuming the NE is unique, Daskalakis and Panageas [2019] showed that OMWU converges in two-player zero-sum NFGs asymptotically. Wei et al. [2021] further improved the result by showing that both OMWU and optimistic gradient descent ascent (OGDA) converge to the NE with a global sublinear convergence rate $O(1/T)$ and a local linear convergence rate in NFGs. Among them, OGDA removes the unique NE assumption.

Last-iterate convergence in EFGs. Some results above, such as Cai et al. [2022], Wei et al. [2021], apply to general convex strategy sets beyond the probability simplex of NFGs. These results can potentially be extended to EFGs. However, their application is limited by using the Euclidean norm as the regularizer, which restricts their extension to efficient updates in EFGs. In other words, they may need to perform quadratic programming at every iteration to project into EFGs' strategy space. On the other hand, to enable efficient closed-form updates of mirror descent, algorithms based on mirror descent typically resort to dilated regularizers [Hoda et al., 2010, Kroer et al., 2020], which poses additional difficulty in showing the last-iterate convergence.

Farina et al. [2019c] first empirically showed that OMD also enjoys last-iterate convergence in EFGs. Then, Lee et al. [2021] extends OMWU from NFGs [Wei et al., 2021] to EFGs, but still requires the unique NE assumption. Piliouras et al. [2022] showed the last-iterate convergence of OGDA in network zero-sum EFGs without the unique NE assumption. However, the regularizer used is neither dilated nor entropy-based, which makes the algorithm less scalable.

Learning with Stochastic Feedback.

External sampling. Lanctot et al. [2009] proposed external sampling in EFGs to estimate the counterfactual values. When updating the strategy of player 1 in a two-player zero-sum

EFG, external sampling will enumerate all actions at the infoset of player 1 and sample an action at other places. In this way, external sampling can obtain an unbiased estimator of the counterfactual value without importance sampling. Previous works about solving Texas hold'em [Brown and Sandholm, 2019b, Brown et al., 2019, Moravčík et al., 2017] mostly depend on external sampling. However, external sampling visits approximately the square root of the number of nodes in the game tree (depending on the game tree structure) at each iteration, which inhibits its deployment in larger games such as Stratego (more than 10^{500} nodes [Perolat et al., 2022]).

Rollout-based estimation (outcome sampling). Lanctot et al. [2009] proposed Outcome-Sampling Monte-Carlo CFR (OS-MCCFR), a variant of CFR which uses random rollouts of trajectories to estimate counterfactual values. Later, Bai et al. [2022], Farina and Sandholm [2021], Farina et al. [2021b], and Fiegel et al. [2023] proposed algorithms that learning in EFGs with rolling trajectories at each iteration. Nonetheless, those algorithms rely on importance sampling, which causes numerical instability due to the large magnitude of the feedback. In large games such as Stratego, the unbiased estimator of counterfactual values can be as large as 10^{200}, which causes numerical instability in training neural networks. In Sokota et al. [2023], they empirically showed last-iterate convergence with rolling trajectories and without importance sampling.

ESCHER [McAleer et al., 2023] and LocalOMD [Fiegel et al., 2024] sample trajectories with a fixed exploration strategy. This is problematic because subgames reached by the fixed exploration strategy may not overlap with those of the current strategy, which will result in insufficient exploration for subgames frequently reached by the current strategy. Furthermore, McAleer et al. [2023] and Fiegel et al. [2024] only guarantee average-iterate convergence.

ARMAC [Gruslys et al., 2020] and ACH [Fu et al., 2021] support both Q-values and approximately-on-policy estimation, but like ESCHER and LocalOMD they do not guarantee convergence in iterate. Moreover, neither of them is computationally efficient since they sample many trajectories (possibly infinite) at each iteration to ensure that the estimation is totally accurate. Plus, ACH does not converge to the set of NEs, even in terms of average-iterate convergence.

Algorithm	Iterate convergence	Q-values	Stochastic feedback
DREAM, DEEP-CFR [Brown et al., 2019, Steinberger et al., 2020]	✗	✗	✓
OOMD, REG-DOMD, MMD [Lee et al., 2021, Liu et al., 2023, Sokota et al., 2023]	✓(last)	✗	✗
ADAPTIVE FTRL [Fiegel et al., 2023]	✗	✗	✓
ESCHER, LOCALOMD [Fiegel et al., 2024, McAleer et al., 2023]	✗	✓	≈ (off-policy)
ARMAC, ACH [Fu et al., 2021, Gruslys et al., 2020]	✗	✓	≈ (infinite samples)
QFR (this thesis)	✓(best)	✓	✓

Table 1.1: The table above is a comparison to related work along three aspects: convergence guarantee, feedback type, and whether sampling (stochastic feedback) is supported or not. (last) and (best) denote last-iterate convergence and best-iterate convergence, respectively.

Contribution

In this thesis, we give a positive answer to the following question:

Is it possible to design a theoretically sound policy gradient method for solving two-player zero-sum extensive-form games that achieves the desiderata **(I)**, **(II)**, **(III)** *listed in the introduction?*

In Chapter 3 and Chapter 4, we propose algorithms that satisfy **(II)** in the full-information setting by adding additional regularization to the objective. We show that both optimistic mirror descent (OMD) and any variants of counterfactual regret minimization (CFR) enjoy last/best-iterate convergence in the game with additional regularization.

In Chapter 5, we will propose the first theoretically sound policy gradient method, QFR, for solving two-player zero-sum EFGs. QFR achieves **(II)** by adding additional regularization as the first part. However, to achieve **(I)** and **(III)** simultaneously, we propose a new regularizer for EFGs called *bidilated regularizer*, which can avoid importance sampling when computing the unbiased estimator from a rolling trajectory. Furthermore, instead of *counterfactual values*, QFR updates the strategy according to *trajectory Q-values* to avoid importance sampling. To ensure provable convergence with trajectory Q-value, we develop a novel learning rate schedule, where the learning rate in each infoset monotonically increases with respect to the infoset's depth.

Chapter 2

Preliminaries

For any vector $\boldsymbol{x} \in \mathbb{R}^n$, let $\|\boldsymbol{x}\|_p$ be its p-norm for any constant $p \geq 0$. By default, let $\|\boldsymbol{x}\|$ be the 2-norm, which is also called the Euclidean norm. We use $\Delta^n \coloneqq \{\boldsymbol{x} \in [0,1]^n \colon \sum_{i=1}^n x_i = 1\}$ to denote the $(n-1)$-dimensional probability simplex. For any convex function $\psi \colon \mathbb{R}^n \to \mathbb{R}$, let $D_\psi(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(0)}) \coloneqq \psi(\boldsymbol{x}^{(1)}) - \psi(\boldsymbol{x}^{(0)}) - \left\langle \nabla\psi(\boldsymbol{x}^{(0)}), \boldsymbol{x}^{(1)} - \boldsymbol{x}^{(0)} \right\rangle$ be its associated Bregman divergence. When ψ is c-strongly convex, $D_\psi(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(0)}) \geq \frac{c}{2} \left\|\boldsymbol{x}^{(1)} - \boldsymbol{x}^{(0)}\right\|^2$ by definition of strongly convexity. For any discrete set $\mathcal{S}$, let $|\mathcal{S}|$ be its cardinality. For any integer $n > 0$, let $[n] \coloneqq \{1, 2, ..., n\}$. Therefore, $|[n]| = n$. For any convex set $\mathcal{C} \subseteq \mathbb{R}^n$, let $\mathrm{Proj}_{\mathcal{C}}(\boldsymbol{x}) = \mathrm{argmin}_{\widehat{\boldsymbol{x}} \in \mathcal{C}} \|\widehat{\boldsymbol{x}} - \boldsymbol{x}\|^2$ be the projection of $\boldsymbol{x} \in \mathbb{R}^n$ to $\mathcal{C}$. In the following, we will introduce the basic notations of extensive-form games (EFGs).

2.1 Game Tree and Information Set

EFGs are played on a rooted game tree, with its root denoted as $\emptyset$. Each node in the game tree is also called *history*. Moreover, at each node, a player will conduct an action, and we call that node *belongs* to this player. Throughout this thesis, we focus on two-player zero-sum EFGs, so that the set of all players is $\{1, 2\} \cup \{c\}$, where c is an artificial player called the *chance player*. Player c controls all the stochastic events in the game sampled from a known distribution, such as dealing cards from a shuffled deck or dice rolling. We use $\mathcal{H}_1, \mathcal{H}_2$ to denote the set of nodes belonging to player $1, 2$, and let $\mathcal{H} \coloneqq \mathcal{H}_1 \cup \mathcal{H}_2$ for notational simplicity. The nodes without any subsequent nodes in the game tree are called *terminal nodes*. Without the loss of generality, due to the tree structure, we assume that each player $i \in [2]$ will receive utility $\mathcal{U}_i(h) \in [-1, 1]$ for any node $h \in \mathcal{H}$, and it is non-zero only if h is a terminal node. Moreover, since the game is zero-sum, $\mathcal{U}_1(h) = -\mathcal{U}_2(h)$.

To model the imperfect information, for each player $i \in [2]$, his node set $\mathcal{H}_i$ are partitioned into *information sets* (or *infosets* for brevity) $s_1, s_2, ..., s_m$, and he cannot distinguish any two nodes from the same infoset. For instance, in a poker game, a player cannot distinguish two nodes only differing in his opponent's private cards. Let $\mathcal{S}_i \coloneqq \{s_1, s_2, ..., s_m\}$ denote the collection of all player i's infosets, and $\mathcal{S} \coloneqq \mathcal{S}_1 \cup \mathcal{S}_2$ for simplicity. Let $p \colon \mathcal{S} \to [2]$ be the function that indicates which player the nodes in an infoset belong to. Furthermore, for all nodes in the same infoset s, they must share the same action set $\mathcal{A}_s$, since player $p(s)$ cannot distinguish them.

A standard assumption in EFG is *perfect recall*, *i.e.,* players can remember all past observations and actions. With perfect recall, for any infoset $s \in \mathcal{S}$, any node $h \in s$ has the same observations along the path from the root $\emptyset$ to h, in the view of player $p(s)$. Due to the tree structure, for any two nodes $h, h' \in \mathcal{H}$, we denote $h \sqsubseteq h'$ if h is the ancestor of h' in the game tree. Moreover, for any infoset $s \in \mathcal{S}$, any action $a \in \mathcal{A}_s$, and node $h \in s$, we denote $(h, a) \sqsubseteq h'$ for a node $h' \in \mathcal{H}$ if the path from the root to h' includes taking action a at h. For any player $i \in [2]$ and infosets $s, s' \in \mathcal{S}_i$, we have $s \sqsubseteq s'$ if there exists $h \in s, h' \in s'$ such that $h \sqsubseteq h'$. Similarly, we can define $(s, a) \sqsubseteq s'$ and $(s, a) \sqsubseteq (s', a')$. Moreover, for any player i and any $s \in \mathcal{S}_i$, the *parent sequence* of s is defined as $\sigma(s) \coloneqq (s', a')$, where $s' \in \mathcal{S}_i, a' \in \mathcal{A}_{s'}$ are the last infoset-action pair that belongs to player i along the path from the root to any node $h \in s$. By the perfect recall assumption, the choice of node $h \in s$ does not affect $\sigma(s)$. When the parent sequence of s does not exist, we denote $\sigma(s) = \emptyset$. Moreover, we can generalize the parent sequence to nodes. For each node $h \in \mathcal{H}$, let $\sigma_i(h) = (s, a)$, where $s \in \mathcal{S}_i, a \in \mathcal{A}_s$, be the last infoset-action pair along the path from the root to h that belongs to player i. Lastly, we define the depth $\mathcal{D}(h)$ of a node $h \in \mathcal{H}$ as the number of nodes minus one along the path from the root to h. Therefore, $\mathcal{D}(\emptyset) = 0$. For any infoset s, we define its depth $\mathcal{D}(s) \coloneqq \max_{h \in s} \mathcal{D}(h)$ as the maximum depth of any node $h \in s$. Moreover, let $\mathcal{D} \coloneqq \max_{h \in \mathcal{H}} \mathcal{D}(h)$ be the maximum depth of the game.

2.2 Behavior-form and Sequence-form Strategies

Due to the imperfect information, any player's strategy should be identical for all nodes in the same infoset, since he cannot distinguish those nodes. Therefore, we can define the *behavior-form strategy* of player i as $\pi_i(a \,|\, s)$ for any $s \in \mathcal{S}_i$ and $a \in \mathcal{A}_s$, *i.e.,* the probability of taking action a at infoset s. For simplicity, let $\boldsymbol{\pi} \coloneqq (\pi_1, \pi_2)$ as the strategy profile of all players. For any player i, given his behavior-form strategy π_i, his *sequence-form representation* [Von Stengel, 1996] can be written as $\mu_i^{\pi_i} \colon \{(s, a)\}_{s \in \mathcal{S}_i, a \in \mathcal{A}_s} \to [0, 1]$, which denotes the probability of reaching infoset s and taking action a contributed by π_i. Formally, we can define it recursively by $\mu_i^{\pi_i}(s, a) = \mu_i^{\pi_i}(\sigma(s))\pi_i(a \,|\, s)$ for any $s \in \mathcal{S}_i, a \in \mathcal{A}_s$, and $\mu_i^{\pi_i}(\emptyset) = 1$. Furthermore, let $\mu_c(h)$ denote the probability of reaching node h contributed by the chance player c. Without loss of generality, we can assume $\mu_c(h) > 0$, otherwise the node can be removed from the game since it will never be reached. For notational simplicity, let $\mu^{\boldsymbol{\pi}} \coloneqq (\mu_1^{\pi_1}, \mu_2^{\pi_2})$ represent the concatenation of the sequence-form strategies $\mu_1^{\pi_1}, \mu_2^{\pi_2}$, and $\mu^{\boldsymbol{\pi}}(s, a) = \mu_{p(s)}^{\pi_{p(s)}}(s, a)$ for any $s \in \mathcal{S}, a \in \mathcal{A}_s$.

When it is clear form the context, we write $\mu_i^{\pi_i} \in \mathbb{R}^{\bigcup_{s \in \mathcal{S}_i} \mathcal{A}_s}$ to represent a vector indexed by infoset-action pairs (s, a), *i.e.,* $(\mu_i^{\pi_i})_{(s,a)} = \mu_i^{\pi_i}(s, a)$. Let $\boldsymbol{U} \in [-1, 1]^{\bigcup_{s \in \mathcal{S}_1} \mathcal{A}_s \times \bigcup_{s \in \mathcal{S}_2} \mathcal{A}_s}$ be the utility matrix of the game, where each element $A_{(s_1, a_1), (s_2, a_2)}$ represents the expected utility of player 2, when player 1 reaches (s_1, a_1) and player 2 reaches (s_2, a_2). Therefore, the expected utility of player 2 can be written as $(\mu_1^{\pi_1})^\top \boldsymbol{U} \mu_2^{\pi_2}$ and that of player 1 is $-(\mu_1^{\pi_1})^\top \boldsymbol{U} \mu_2^{\pi_2}$ since the game is zero-sum. Let Π_i be the convex polytope of $\mu_i^{\pi_i}$ and $\Pi \coloneqq \Pi_1 \times \Pi_2$. Then, the objective of the game is solving

$$\min_{\mu_1^{\pi_1} \in \Pi_1} \max_{\mu_2^{\pi_2} \in \Pi_2} (\mu_1^{\pi_1})^\top \boldsymbol{U} \mu_2^{\pi_2}. \tag{2.2.1}$$

A solution (solution may not be unique) $\mu^* = (\mu_1^*, \mu_2^*)$ of (2.2.1) is called *Nash equilibrium* (NE) of the game.

2.3 Counterfactual and Q-Values

This section recalls some basic notions in EFGs, including counterfactual values, Q-values, and trajectory Q-values, which are different estimations of the expected value of taking action a at infoset s. For ease of presentation, in the following, we only consider the values of player 1, since two players are symmetric.

2.3.1 Trajectory Q-Values

For any infoset $s \in \mathcal{S}_1$ and action $a \in \mathcal{A}_s$, its trajectory Q-value $\overline{Q}_1^{\boldsymbol{\pi}}(s,a)$ associated with $\boldsymbol{\pi} = (\pi_1, \pi_2)$, is player 1's expected utility of taking a at s, when players follow the strategy π_1, π_2 respectively. Formally,

$$\overline{Q}_1^{\boldsymbol{\pi}}(s,a) \coloneqq \frac{1}{\pi_1(a \mid s)} \sum_{h' \colon \exists h \in s, (h,a) \sqsubseteq h'} \mu_c(h') \mu_1^{\pi_1}(\sigma_1(h')) \mu_2^{\pi_2}(\sigma_2(h')) \mathcal{U}_1(h'). \qquad \text{(Trajectory Q)}$$

2.3.2 Q-Values

The definition of Q-value in EFGs aligns with that in reinforcement learning literature [Sutton and Barto, 2018]. Specifically, for any infoset $s \in \mathcal{S}_1$ and action $a \in \mathcal{A}_s$, its Q-value $Q_1^{\boldsymbol{\pi}}(s,a)$ is player 1's expected utility of taking a at s, when players follow the strategy π_1, π_2 respectively, conditioned on reaching infoset s and taking action a. Formally,

$$Q_1^{\boldsymbol{\pi}}(s,a) \coloneqq \frac{\overline{Q}_1^{\boldsymbol{\pi}}(s,a))}{\sum_{h \in s} \mu_c(h) \mu_1^{\pi_1}(\sigma_1(h)) \mu_2^{\pi_2}(\sigma_2(h))}. \qquad \text{(Q)}$$

2.3.3 Counterfactual Values

For any infoset $s \in \mathcal{S}_1$ and action $a \in \mathcal{A}_s$, its counterfactual value is player 1's expected utility of taking a at s, when players follow the strategy π_1, π_2 respectively, conditioned on only player 1 reaching infoset s and taking action a. Formally,

$$\mathrm{CF}_1^{\boldsymbol{\pi}}(s,a) \coloneqq \frac{1}{\pi_1(a \mid s)} \sum_{h' \colon \exists h \in s, (h,a) \sqsubseteq h'} \mu_c(h') \frac{\mu_1^{\pi_1}(\sigma_1(h'))}{\mu_1^{\pi_1}(s,a)} \mu_2^{\pi_2}(\sigma_2(h')) \mathcal{U}_1(h'). \qquad \text{(Counterfactual)}$$

Unlike (Trajectory Q) and (Q), we can see that the contributions of players 1 and 2 are asymmetric in (Counterfactual). Another interpretation of counterfactual value is that player 1 plays deterministically to (s,a) and then follows π_1, while player 2 follows π_2 all along the game.

2.4 Convergence Guarantees

In this section, we will define three common convergence types of an iterative algorithm for solving the equilibrium in games. Suppose the algorithm generates strategies $\mu^{(1)}, \mu^{(2)}, ..., \mu^{(T)}$ in $T > 0$ iterations.

Definition 2.4.1 (Average-Iterate Convergence)**.** An algorithm enjoys average-iterate convergence if

$$\frac{1}{T}\sum_{t=1}^{T}\mu^{(t)}$$

converges to the equilibrium.

We will introduce two adverse operations before introducing best-iterate convergence and last-iterate convergence, since they will trivialize the convergence guarantees if allowed.

Definition 2.4.2 (Adverse Operations)**.** We call the following two operations in an algorithm an adverse operation,

(i) Taking a linear combination of iterates during the execution of the algorithm.

(ii) Computing the exploitability or other measures on the distance to the equilibrium.

Then, we will introduce the best-iterate convergence.

Definition 2.4.3 (Best-Iterate Convergence)**.** An algorithm enjoys best-iterate convergence if

$$\exists t^* \in [T],\ \mu^{t^*}$$

converges to the equilibrium, *without adverse operation* (i).

Suppose (i) is allowed. In that case, we can convert any algorithm satisfying average-iterate convergence to best-iterate convergence by maintaining $\mu^{\text{avg},(t+1)} := \frac{t}{t+1}\mu^{\text{avg},(t)} + \frac{1}{t+1}\mu^{(t)} = \frac{1}{t+1}\sum_{s=1}^{t+1}\mu^{(s)}$. Then, once $\mu^{(1)}, \mu^{(2)}, ..., \mu^{(T)}$ satisfies average-iterate convergence, $\mu^{\text{avg},(1)}, \mu^{\text{avg},(2)}, ..., \mu^{\text{avg},(T)}$ satisfies best-iterate convergence (and even last-iterate convergence introduced below).

Definition 2.4.4 (Last-Iterate Convergence)**.** An algorithm enjoys last-iterate convergence if

$$\mu^{(T)}$$

converges to the equilibrium, *without adverse operation* $(i), (ii)$.

(i) cannot be allowed according to the same reason as best-iterate convergence. (ii) cannot be allowed because otherwise, any algorithm with best-iterate convergence will also enjoy last-iterate convergence, by maintaining

$$\mu^{\text{best},(t+1)} := \begin{cases} \mu^{\text{best},(t)} & \text{dist}(\mu^*, \mu^{\text{best},(t)}) \leq \text{dist}(\mu^*, \mu^{(t+1)}) \\ \mu^{(t+1)} & \text{Otherwise.} \end{cases}$$

Then, once $\mu^{(1)}, \mu^{(2)}, ..., \mu^{(T)}$ satisfies best-iterate convergence, $\mu^{\text{best},(1)}, \mu^{\text{best},(2)}, ..., \mu^{\text{best},(T)}$ satisfies last-iterate convergence.

Moreover, we call best-iterate convergence and last-iterate convergence as *iterate convergence*.

2.5 (Bi)Dilated Regularizer

In this section, we will introduce regularizers in EFGs, including *dilated regularizer* and *bidilated regularizer*. In two-player zero-sum games, adding an additional strongly convex functions (regularizers) to the original bilinear objective will make it strongly convex with respect to each player's strategy. Then, we may derive algorithms with faster convergence rates or more desirable convergence properties, such as best-iterate and last-iterate convergence.

An intuitive regularizer for EFGs is $\psi^{\Pi}_{\text{vanilla}} := \|\mu_1^{\pi_1}\|^2 + \|\mu_2^{\pi_2}\|^2$ for sequence-form strategy profile $\mu^{\boldsymbol{\pi}}$. However, there is no closed-form solution to the projection into Π with respect to the Bregman divergence associated with $\psi^{\Pi}_{\text{vanilla}}$ [Lee et al., 2021], *i.e.,* $\operatorname{argmin}_{\mu^{\boldsymbol{\pi}} \in \Pi} D_{\psi^{\Pi}_{\text{vanilla}}}(\mu^{\boldsymbol{\pi}}, \mu^{\bar{\boldsymbol{\pi}}})$ for arbitrary sequence-form strategy profile $\mu^{\bar{\boldsymbol{\pi}}} \in \Pi$. Such limitation hinders applying $\psi^{\Pi}_{\text{vanilla}}$ to mirror descent in EFGs. Therefore, in EFGs, dilated regularizer [Hoda et al., 2010] and bidilated regularizer are more suitable since there are closed-form solutions to the projection with respect to the Bregman divergence associated with the (bi)dilated regularizers.

2.5.1 Dilated Regularizer

For each infoset $s \in \mathcal{S}$, we can define its local regularizer as

$$\psi_s^{\Delta}(\pi_{p(s)}) = \begin{cases} \frac{\alpha_s}{2} \sum_{a \in \mathcal{A}_s} \left(\pi_{p(s)}(a \mid s)\right)^2 & \text{(Euclidean Norm)} \\ \alpha_s \left(\log |\mathcal{A}_s| + \sum_{a \in \mathcal{A}_s} \pi_{p(s)}(a \mid s) \log \pi_{p(s)}(a \mid s)\right) & \text{(Negative Entropy),} \end{cases} \tag{2.5.1}$$

where α_s is a hyper-parameter depending on the infoset s. Typically, we choose Euclidean norm and negative entropy as the regularizer, but any strongly convex function works here. Then, the dilated regularizer for player $i \in [2]$'s strategy is the sum of local regularizers over all infosets $s \in \mathcal{S}_i$ and weighted by the reach probability of player i to that infoset. Formally,

$$\psi^{\Pi_i}(\mu_i^{\pi_i}) := \sum_{s \in \mathcal{S}_i} \mu_i^{\pi_i}(\sigma(s)) \psi_s^{\Delta}(\pi_i). \tag{2.5.2}$$

By choosing α_s properly, $\psi^{\Pi_i}(\mu_i^{\pi_i})$ can be 1-strongly convex with respect to 2-norm, *i.e.,* for any sequence-form strategies $\mu_i^{\pi_i^{(1)}}, \mu_i^{\pi_i^{(2)}} \in \Pi_i$,

$$D_{\psi^{\Pi_i}}(\mu_i^{\pi_i^{(2)}}, \mu_i^{\pi_i^{(1)}}) \geq \left\| \mu_i^{\pi_i^{(2)}} - \mu_i^{\pi_i^{(1)}} \right\|^2.$$

Throughout this thesis, we assume ψ^{Π} is 1-strongly convex with respect to 2-norm. For notational simplicity, let $\psi^{\Pi}(\mu^{\boldsymbol{\pi}}) := \psi^{\Pi_1}(\mu_1^{\pi_1}) + \psi^{\Pi_2}(\mu_2^{\pi_2})$.

2.5.2 Bidilated Regularizer

When learning from trajectory feedback, dilated regularizer in (2.5.2) requires importance sampling, since it is weighted only by the reach probability of player i (see Chapter 5 for details). Therefore, we propose bidilated regularizer where the local regularizer in each infoset is weighted by the reach probability contributed by player 1, player 2, and the chance player. Formally, the bidilated regularizer for player 1 (similarly for that of player 2) can be written as,

$$\psi_{\mathrm{bi}}^{\Pi_1}(\mu_1^{\pi_1}, \mu_2^{\pi_2}) \coloneqq \sum_{s \in \mathcal{S}_1} \mu_1^{\pi_1}(\sigma(s)) \left(\sum_{h \in s} \mu_c(h) \mu_2^{\pi_2}(\sigma_2(h)) \right) \psi_s^{\Delta}(\pi_1). \tag{2.5.3}$$

For notational simplicity, let $\psi_{\mathrm{bi}}^{\Pi}(\mu^{\boldsymbol{\pi}}) \coloneqq \psi_{\mathrm{bi}}^{\Pi_1}(\mu_1^{\pi_1}, \mu_2^{\pi_2}) + \psi_{\mathrm{bi}}^{\Pi_2}(\mu_1^{\pi_1}, \mu_2^{\pi_2})$.

2.5.3 Regularized Game

By adding (bi)dilated regularizer to the original objective (2.2.1), we have the following objectives, (2.5.4) and (2.5.5), which are called the *regularized game*. In contrast, (2.2.1) is called the *original game*.

$$\min_{\mu_1^{\pi_1} \in \Pi_1} \max_{\mu_2^{\pi_2} \in \Pi_2} (\mu_1^{\pi_1})^\top \boldsymbol{U} \mu_2^{\pi_2} + \tau \psi^{\Pi_1}(\mu_1^{\pi_1}) - \tau \psi^{\Pi_2}(\mu_2^{\pi_2}) \tag{2.5.4}$$

$$\min_{\mu_1^{\pi_1} \in \Pi_1} \max_{\mu_2^{\pi_2} \in \Pi_2} (\mu_1^{\pi_1})^\top \boldsymbol{U} \mu_2^{\pi_2} + \tau \psi_{\mathrm{bi}}^{\Pi_1}(\mu_1^{\pi_1}, \mu_2^{\pi_2}) - \tau \psi_{\mathrm{bi}}^{\Pi_2}(\mu_1^{\pi_1}, \mu_2^{\pi_2}). \tag{2.5.5}$$

Furthermore, the solution of the regularized game is called the *regularized Nash equilibrium*, which is unique when $\tau > 0$ due to the strong convexity. We denote it as $\mu^{\tau,*}$, and whether it is the solution to (2.5.4) or (2.5.5) depends on the context.

2.6 Regret and Equilibrium

No-regret learning is one of the key tools for solving the NE. Specifically, let $\pi_i^{(t)}$ be the strategy of player $i \in [2]$ at timestep t and $\mu_i^{(t)} \coloneqq \mu_i^{\pi_i^{(t)}}$ be the short-hand for player i's sequence-form strategy, then we can define player 1's regret of the game as

$$G_1^{(T)}(\mu_1) \coloneqq \sum_{t=1}^{T} \left\langle \boldsymbol{U} \mu_2^{(t)}, \mu_1^{(t)} - \mu_1 \right\rangle \tag{Difference}$$

$$R_1^{(T)} \coloneqq \max_{\widehat{\mu}_1 \in \Pi_1} G_1^{(T)}(\widehat{\mu}_1). \tag{Regret}$$

$G_2^{(T)}(\mu_2), R_2^{(T)}$ can be defined similarly. The Folk theorem shows that when both $R_1^{(T)}$ and $R_2^{(T)}$ are sub-linear in T, then $\left(\mu_1^{(t)}, \mu_2^{(t)}\right)_{t=1}^{T}$ enjoys average-iterate convergence to the NE. Formally, let $\bar{\mu}_i \coloneqq \frac{1}{T} \sum_{t=1}^{T} \mu_i^{(t)}$ be the average sequence-form strategy of player i, we have

$$\max_{\widehat{\mu}_1 \in \Pi_1} \langle \boldsymbol{U} \bar{\mu}_2, \bar{\mu}_1 - \widehat{\mu}_1 \rangle + \max_{\widehat{\mu}_2 \in \Pi_2} \langle -\boldsymbol{U}^\top \bar{\mu}_1, \bar{\mu}_2 - \widehat{\mu}_2 \rangle \leq \frac{R_1^{(T)} + R_2^{(T)}}{T}. \tag{Exploitability}$$

The left-hand side of the equation above is called *exploitability* of $(\bar{\mu}_1, \bar{\mu}_2)$. For any strategy profile $\boldsymbol{\mu} \coloneqq (\mu_1, \mu_2)$ that the exploitability is bounded by ϵ, we call it ϵ-approximate NE. Furthermore, when we imposed additional regularization to the objective function, such as (2.5.4) and (2.5.5), we can also define the difference and regret correspondingly as

$$G_1^{\psi,(T)}(\mu_1) \coloneqq \sum_{t=1}^{T} \left(\left\langle \boldsymbol{U} \mu_2^{(t)}, \mu_1^{(t)} - \mu_1 \right\rangle + \tau\psi(\mu_1^{(t)}) - \tau\psi\left(\mu_1\right) \right) \tag{Difference}$$

$$R_1^{\psi,(T)} \coloneqq \max_{\widehat{\mu}_1 \in \Pi_1} G_1^{\psi,(T)}(\widehat{\mu}_1), \tag{Regret}$$

where ψ can be either ψ^{Π_1} or $\psi_{\mathrm{bi}}^{\Pi_1}$.

Chapter 3

Regularized Dilated Optimistic Mirror Descent (Reg-DOMD)

We slightly modify the Regularized Dilated Optimistic Mirror Descent (`Reg-DOMD`) proposed in Liu et al. [2023]. In the following, let $\mu^{(t)} := \mu^{\boldsymbol{\pi}^{(t)}}$ and $\bar{\mu}^{(t)} := \mu^{\bar{\boldsymbol{\pi}}^{(t)}}$ for notational simplicity. Moreover, the gradient of sequence-form strategy $\mu^{\boldsymbol{\pi}}$ is defined as $F(\mu^{\boldsymbol{\pi}}) := \left(\boldsymbol{U}\mu_2^{\pi_2}, -\boldsymbol{U}^\top \mu_1^{\pi_1}\right)$.

Proposition 3.0.1. Function $F(\mu^{\boldsymbol{\pi}})$ is L-Lipschitz continuous. Formally, for any two strategy profiles $\boldsymbol{\pi}, \boldsymbol{\pi}'$, we have

$$\left\| F(\mu^{\boldsymbol{\pi}}) - F(\mu^{\boldsymbol{\pi}'}) \right\| \leq L \left\| \mu^{\boldsymbol{\pi}} - \mu^{\boldsymbol{\pi}'} \right\|. \tag{3.0.1}$$

Since every element in the utility matrix $\boldsymbol{U}$ Proposition 3.0.1 is bounded in $[-1, 1]$, we can easily show that $L \leq 2\sqrt{\left|\bigcup_{s\in\mathcal{S}_1} \mathcal{A}_s\right| \cdot \left|\bigcup_{s\in\mathcal{S}_2} \mathcal{A}_s\right|}$, which is further bounded by twice the number of nodes in the game tree [1].

Then, for any player $i \in [2]$, and timestep $t \in [T]$, the update-rule of `Reg-DOMD` is

$$\begin{aligned} \mu^{(t)} &= \operatorname*{argmin}_{\mu^{\boldsymbol{\pi}}\in\Pi} \left\langle \mu^{\boldsymbol{\pi}}, F(\mu^{(t-1)}) \right\rangle + \tau\psi^{\Pi}(\mu^{\boldsymbol{\pi}}) + \frac{1}{\eta} D_{\psi^{\Pi}}\left(\mu^{\boldsymbol{\pi}}, \bar{\mu}^{(t)}\right) \\ \bar{\mu}^{(t+1)} &= \operatorname*{argmin}_{\mu^{\boldsymbol{\pi}}\in\Pi} \left\langle \mu^{\boldsymbol{\pi}}, F(\mu^{(t)}) \right\rangle + \tau\psi^{\Pi}(\mu^{\boldsymbol{\pi}}) + \frac{1}{\eta} D_{\psi^{\Pi}}\left(\mu^{\boldsymbol{\pi}}, \bar{\mu}^{(t)}\right), \end{aligned} \tag{Reg-DOMD}$$

where $\eta > 0, \tau \geq 0$ are hyper-parameters that control the learning rate and regularization respectively. Note that $\boldsymbol{\pi}^{(0)}, \bar{\boldsymbol{\pi}}^{(1)}$ are initialized as uniform strategy over $\mathcal{A}_s$ in every infoset $s \in \mathcal{S}$. (Reg-DOMD) is called Regularized Dilated Optimistic Multiplicative Weights Update (`Reg-DOMWU`) when ψ^{Π} is negative entropy, and Regularized Dilated Optimistic Gradient Descent Ascent (`Reg-DOGDA`) when ψ^{Π} is Euclidean norm.

In the following, we show that `Reg-DOMD` enjoys linear convergence to the equilibrium $\mu^{\tau,*}$ of the regularized objective, (2.5.4).

This chapter builds on Section 4 of Liu et al. [2023].

[1] In practice, L may be much smaller than this upperbound.

Theorem 3.0.2. Consider (Reg-DOMD). When $\eta \le \frac{1}{2L}$, $\tau \le 1$, and ψ^{Π} being a 1-strongly convex function, (Reg-DOMD) guarantees that

$$D_{\psi^{\Pi}}\left(\mu^{\tau,*}, \bar{\mu}^{(t+1)}\right) \le \left(\frac{1}{1+\eta\tau}\right)^{t} D_{\psi^{\Pi}}\left(\mu^{\tau,*}, \bar{\mu}^{(1)}\right), \tag{3.0.2}$$

for any $t \ge 1$.

By analyzing (Reg-DOMD), we can get the following lemma by the three-point identity of Bregman divergence.

Lemma 3.0.3. Let $\mathcal{C}$ be a convex set and $\boldsymbol{x}^{(1)} = \operatorname{argmin}_{\boldsymbol{x}\in\mathcal{C}} \left\{ \langle \boldsymbol{g}, \boldsymbol{x}\rangle + \tau_0 \psi^{\mathcal{C}}(\boldsymbol{x}) + \frac{1}{\eta} D_{\psi^{\mathcal{C}}}(\boldsymbol{x}, \boldsymbol{x}^{(0)}) \right\}$, where $\psi^{\mathcal{C}}$ is a strongly-convex function in $\mathcal{C}$ and $\tau_0 \ge 0$ is a constant. Then, for any $\boldsymbol{x}^{(2)} \in \mathcal{C}$, we have

$$\begin{aligned} &\eta\tau_0\psi^{\mathcal{C}}(\boldsymbol{x}^{(1)}) - \eta\tau_0\psi^{\mathcal{C}}(\boldsymbol{x}^{(2)}) + \eta\left\langle \boldsymbol{g}, \boldsymbol{x}^{(1)} - \boldsymbol{x}^{(2)}\right\rangle \\ \le& D_{\psi^{\mathcal{C}}}(\boldsymbol{x}^{(2)}, \boldsymbol{x}^{(0)}) - (1+\eta\tau_0) D_{\psi^{\mathcal{C}}}(\boldsymbol{x}^{(2)}, \boldsymbol{x}^{(1)}) - D_{\psi^{\mathcal{C}}}(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(0)}). \end{aligned} \tag{3.0.3}$$

The proof is postponed to Appendix A.

By plugging $\boldsymbol{x}^{(0)} = \bar{\mu}^{(t)}, \boldsymbol{x}^{(1)} = \mu^{(t)}, \boldsymbol{g} = F(\mu^{(t-1)}), \tau_0 = \tau, \psi^{\mathcal{C}} = \psi^{\Pi}$, and $\boldsymbol{x}^{(2)} = \bar{\mu}^{(t+1)}$, into Lemma C.3.4, we have

$$\begin{aligned} &\eta\tau\psi^{\Pi}(\mu^{(t)}) - \eta\tau\psi^{\Pi}(\bar{\mu}^{(t+1)}) + \eta\left\langle F(\mu^{(t-1)}), \mu^{(t)} - \bar{\mu}^{(t+1)}\right\rangle \\ \le& D_{\psi^{\Pi}}(\bar{\mu}^{(t+1)}, \bar{\mu}^{(t)}) - (1+\eta\tau) D_{\psi^{\Pi}}(\bar{\mu}^{(t+1)}, \mu^{(t)}) - D_{\psi^{\Pi}}(\mu^{(t)}, \bar{\mu}^{(t)}). \end{aligned}$$

Similarly, by plugging $\boldsymbol{x}^{(0)} = \bar{\mu}^{(t)}, \boldsymbol{x}^{(1)} = \bar{\mu}^{(t+1)}, \boldsymbol{g} = F(\mu^{(t)}), \tau_0 = \tau, \psi^{\mathcal{C}} = \psi^{\Pi}$, and $\boldsymbol{x}^{(2)} = \mu$ where μ is an arbitrary sequence-form strategy profile, into Lemma C.3.4, we have

$$\begin{aligned} &\eta\tau\psi^{\Pi}(\bar{\mu}^{(t+1)}) - \eta\tau\psi^{\Pi}(\mu) + \eta\left\langle F(\mu^{(t)}), \bar{\mu}^{(t+1)} - \mu\right\rangle \\ \le& D_{\psi^{\Pi}}(\mu, \bar{\mu}^{(t)}) - (1+\eta\tau) D_{\psi^{\Pi}}(\mu, \bar{\mu}^{(t+1)}) - D_{\psi^{\Pi}}(\bar{\mu}^{(t+1)}, \bar{\mu}^{(t)}). \end{aligned}$$

By summing them up and adding $\eta\left\langle F(\mu^{(t)}) - F(\mu^{(t-1)}), \mu^{(t)} - \bar{\mu}^{(t+1)}\right\rangle$, we have

$$\begin{aligned} &\eta\tau\psi^{\Pi}(\mu^{(t)}) - \eta\tau\psi^{\Pi}(\mu) + \eta\left\langle F(\mu^{(t)}), \mu^{(t)} - \mu\right\rangle \\ \le& -(1+\eta\tau) D_{\psi^{\Pi}}(\bar{\mu}^{(t+1)}, \mu^{(t)}) - D_{\psi^{\Pi}}(\mu^{(t)}, \bar{\mu}^{(t)}) + D_{\psi^{\Pi}}(\mu, \bar{\mu}^{(t)}) - (1+\eta\tau) D_{\psi^{\Pi}}(\mu, \bar{\mu}^{(t+1)}) \\ &+ \eta\left\langle F(\mu^{(t)}) - F(\mu^{(t-1)}), \mu^{(t)} - \bar{\mu}^{(t+1)}\right\rangle \end{aligned}$$

To further bound $\eta\left\langle F(\mu^{(t)}) - F(\mu^{(t-1)}), \mu^{(t)} - \bar{\mu}^{(t+1)}\right\rangle$, we invoke the following lemma.

Lemma 3.0.4. Suppose that $\psi^{\mathcal{C}}$ is a 1-strongly convex function with respect to p-norm in convex set $\mathcal{C}$, such that $D_{\psi^{\mathcal{C}}}(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(0)}) \ge \frac{1}{2}\|\boldsymbol{x}^{(1)} - \boldsymbol{x}^{(0)}\|_p^2$ for any $\boldsymbol{x}^{(0)}, \boldsymbol{x}^{(1)} \in \mathcal{C}$. Then, for any $\boldsymbol{x}^{(0)}, \boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)} \in \mathcal{C}$ and $\tau_0 \ge 0$ satisfying,

$$\begin{aligned} \boldsymbol{x}^{(1)} &= \operatorname*{argmin}_{\boldsymbol{x}\in\mathcal{C}} \left\langle \boldsymbol{x}, \boldsymbol{g}^{(1)}\right\rangle + \tau_0 \nabla\psi^{\mathcal{C}}(\boldsymbol{x}) + D_{\psi^{\mathcal{C}}}(\boldsymbol{x}, \boldsymbol{x}^{(0)}) \\ \boldsymbol{x}^{(2)} &= \operatorname*{argmin}_{\boldsymbol{x}\in\mathcal{C}} \left\langle \boldsymbol{x}, \boldsymbol{g}^{(2)}\right\rangle + \tau_0 \nabla\psi^{\mathcal{C}}(\boldsymbol{x}) + D_{\psi^{\mathcal{C}}}(\boldsymbol{x}, \boldsymbol{x}^{(0)}), \end{aligned}$$

we have

$$\left\|\boldsymbol{x}^{(1)}-\boldsymbol{x}^{(2)}\right\|_p \leq \frac{1}{1+\tau_0}\left\|\boldsymbol{g}^{(1)}-\boldsymbol{g}^{(2)}\right\|_q, \tag{3.0.4}$$

where $q \geq 1$ and $\frac{1}{p}+\frac{1}{q}=1$.

The proof is postponed to Appendix A. Then, by update-rule (Reg-DOMD) and Lemma 3.0.4, we have

$$\begin{aligned}\eta\left\langle F(\mu^{(t)})-F(\mu^{(t-1)}),\mu^{(t)}-\bar{\mu}^{(t+1)}\right\rangle \overset{(i)}{\leq}&\eta\left\|F(\mu^{(t)})-F(\mu^{(t-1)})\right\|\cdot\left\|\mu^{(t)}-\bar{\mu}^{(t+1)}\right\|\\ \leq&\eta^2\left\|F(\mu^{(t)})-F(\mu^{(t-1)})\right\|^2\\ \overset{(ii)}{\leq}&\eta^2L^2\left\|\mu^{(t)}-\mu^{(t-1)}\right\|^2.\end{aligned}$$

(i) is by Hölder's Inequality. (ii) uses Proposition 3.0.1. When we pick $\eta \leq \frac{1}{2L}$, then,

$$\begin{aligned}\eta\left\langle F(\mu^{(t)})-F(\mu^{(t-1)}),\mu^{(t)}-\bar{\mu}^{(t+1)}\right\rangle \leq&\frac{1}{4}\left\|\mu^{(t)}-\mu^{(t-1)}\right\|^2\\ \overset{(i)}{\leq}&\frac{1}{2}\left\|\mu^{(t)}-\bar{\mu}^{(t)}\right\|^2+\frac{1}{2}\left\|\bar{\mu}^{(t)}-\mu^{(t-1)}\right\|^2\\ \overset{(ii)}{\leq}&D_{\psi^\Pi}\left(\mu^{(t)},\bar{\mu}^{(t)}\right)+D_{\psi^\Pi}\left(\bar{\mu}^{(t)},\mu^{(t-1)}\right).\end{aligned}$$

(i) is because $(a+b)^2 \leq 2a^2+2b^2$ for any $a,b\in\mathbb{R}$. (ii) is by 1-strong convexity of ψ^Π. Therefore,

$$\begin{aligned}&\eta\tau\psi^\Pi(\mu^{(t)})-\eta\tau\psi^\Pi(\mu)+\eta\left\langle F(\mu^{(t)}),\mu^{(t)}-\mu\right\rangle\\ \leq&-(1+\eta\tau)D_{\psi^\Pi}(\bar{\mu}^{(t+1)},\mu^{(t)})-D_{\psi^\Pi}(\mu^{(t)},\bar{\mu}^{(t)})+D_{\psi^\Pi}(\mu,\bar{\mu}^{(t)})-(1+\eta\tau)D_{\psi^\Pi}(\mu,\bar{\mu}^{(t+1)})\\ &+D_{\psi^\Pi}\left(\mu^{(t)},\bar{\mu}^{(t)}\right)+D_{\psi^\Pi}\left(\bar{\mu}^{(t)},\mu^{(t-1)}\right)\\ =&D_{\psi^\Pi}(\mu,\bar{\mu}^{(t)})-(1+\eta\tau)D_{\psi^\Pi}(\mu,\bar{\mu}^{(t+1)})+D_{\psi^\Pi}\left(\bar{\mu}^{(t)},\mu^{(t-1)}\right)-(1+\eta\tau)D_{\psi^\Pi}(\bar{\mu}^{(t+1)},\mu^{(t)}).\end{aligned}$$

By the following lemma, the left-hand side of the inequality above can be lowerbounded when $\mu=\mu^{\tau,*}$, the NE of the regularized objective (2.5.4).

Lemma 3.0.5. For any $\tau_0 \geq 0$ and sequence-form strategy profile μ, we have

$$\tau_0\psi^\Pi(\mu)-\tau_0\psi^\Pi(\mu^{\tau_0,*})+\left\langle F(\mu),\mu-\mu^{\tau_0,*}\right\rangle \geq \tau_0 D_{\psi^\Pi}(\mu,\mu^{\tau_0,*}). \tag{3.0.5}$$

The proof is postponed to Appendix A. Then,

$$\begin{aligned}&D_{\psi^\Pi}(\mu^{\tau_0,*},\bar{\mu}^{(t)})-(1+\eta\tau)D_{\psi^\Pi}(\mu^{\tau_0,*},\bar{\mu}^{(t+1)})+D_{\psi^\Pi}\left(\bar{\mu}^{(t)},\mu^{(t-1)}\right)-(1+\eta\tau)D_{\psi^\Pi}(\bar{\mu}^{(t+1)},\mu^{(t)})\\ \geq&\eta\tau D_{\psi^\Pi}(\mu,\mu^{\tau_0,*})\overset{(i)}{\geq}0.\end{aligned}$$

(i) uses the non-negativity of Bregman divergence. By letting $\Theta^{(t+1)} \coloneqq D_{\psi^\Pi}(\mu^{\tau_0,*},\bar{\mu}^{(t+1)})+D_{\psi^\Pi}\left(\bar{\mu}^{(t+1)},\mu^{(t)}\right)$ and rearranging the terms, the inequality above can be written as,

$$\Theta^{(t+1)} \leq \frac{1}{1+\eta\tau}\Theta^{(t)}.$$

Finally,

$$D_{\psi^\Pi}\left(\mu^{\tau,*}, \bar{\mu}^{(t+1)}\right) \leq \Theta^{(t+1)} \leq \left(\frac{1}{1+\eta\tau}\right)^t \Theta^{(1)} \overset{(i)}{=} \left(\frac{1}{1+\eta\tau}\right)^t D_{\psi^\Pi}\left(\mu^{\tau,*}, \bar{\mu}^{(1)}\right).$$

(i) uses the fact that $\bar{\mu}^{(1)} = \mu^{(0)}$ by the initialization of (Reg-DOMD). □

3.1 From the Regularized Game to the Original Game

Intuitively, if the weight of regularization τ is sufficiently small, the regularized NE should be close to the NE of the original game, since the objective functions only differ by $\mathcal{O}(\tau)$. Formally,

Lemma 3.1.1. For any $\tau > 0$ and sequence-form strategy profile $\mu^{\boldsymbol{\pi}} \in \Pi$, we have

$$\max_{\mu^{\widehat{\boldsymbol{\pi}}} \in \Pi} \left\langle F(\mu^{\boldsymbol{\pi}}), \mu^{\boldsymbol{\pi}} - \mu^{\widehat{\boldsymbol{\pi}}} \right\rangle \leq \tau \max_{\mu^{\boldsymbol{\pi}'} \in \Pi} \psi^\Pi(\mu^{\boldsymbol{\pi}'}) + \sqrt{\left(\sum_{s \in \mathcal{S}} |\mathcal{A}_s|\right) D_{\psi^\Pi}(\mu^{\tau,*}, \mu^{\boldsymbol{\pi}})}. \tag{3.1.1}$$

The proof is postponed to Appendix A.1.

In the following, we show how Theorem 3.0.2 implies a last-iterate guarantee to the original NE.

We shrink the weight of regularization τ as follows: (i). Initialize $\tau = \tau_0$ for some hyper-parameter τ_0 at the beginning and run (Reg-DOMD) for several episodes. In each episode, we update the strategies by (Reg-DOMD) for $\widetilde{\Theta}(1/\tau)$ iterations. Then, the duality gap of $\bar{\mu}^{(t)}$ will be no larger than $\mathcal{O}(\tau)$ according to Lemma 3.1.1 and Theorem 3.0.2. Then, we will shrink τ by one-half and start the next episode from scratch.

Theorem 3.1.2. With the shrinking algorithm described above, the duality gap satisfies $\max_{\mu^{\widehat{\boldsymbol{\pi}}} \in \Pi} \left\langle F(\bar{\mu}^{(t+1)}, \bar{\mu}^{(t+1)} - \mu^{\widehat{\boldsymbol{\pi}}} \right\rangle \leq \widetilde{\mathcal{O}}\left(\frac{1}{t}\right)$ for $t = 1, 2, ..., T$. Moreover, let Π^* be the set of NE of the original game, the distance to the set of NE also satisfies $\left\|\bar{\mu}^{(t+1)} - \mathrm{Proj}_{\Pi^*}\left(\bar{\mu}^{(t+1)}\right)\right\| \leq \widetilde{\mathcal{O}}\left(\frac{1}{t}\right)$.

We briefly sketch the intuition behind the proof in the following and defer the full details to Appendix A.1.

Technical Overview. According to Lemma 3.1.1 and Theorem 3.0.2, it is straightforward to prove the first half of Theorem 3.1.2.

Nonetheless, this argument does not imply a small distance to Π^*, because the distance between $\mu^{\tau,*}$ and μ^* is unknown. Instead, we need to show the "slope" of the duality gap is strictly positive, *i.e.*,, for any $\mu^{\boldsymbol{\pi}}$, we have $\max_{\mu^{\widehat{\boldsymbol{\pi}}} \in \Pi} \left\langle F(\mu^{\boldsymbol{\pi}}), \mu^{\boldsymbol{\pi}} - \mu^{\widehat{\boldsymbol{\pi}}} \right\rangle \geq c \left\|\bar{\mu}^{\boldsymbol{\pi}} - \mathrm{Proj}_{\Pi^*}\left(\bar{\mu}^{\boldsymbol{\pi}}\right)\right\|$ for some constant $c > 0$. A proof about the positivity of the "slope" can be found in Gilpin et al. [2008], Liu et al. [2023], Wei et al. [2021]. However, the slope c depends on the utility matrix $\boldsymbol{U}$ and can be arbitrarily small.

Chapter 4

Regularized Counterfactual Regret Minimization (Reg-CFR)

In this chapter, we will propose a generalized version of the `Reg-CFR` in Liu et al. [2023]. Specifically, it is known that by minimizing the local regret associated with the counterfactual values in each infoset, the global regret will also be minimized [Zinkevich et al., 2007] and thus the average sequence-form strategy will converge to NE. This is called *regret decomposition*.

However, we will further show that by imposing additional regularization to the utility function, any algorithm that enjoys average-iterate convergence by regret decomposition will further enjoy best-iterate convergence to the NE of the regularized game, *i.e.,* $\mu^{\tau,*}$. Recall that best-iterate (Definition 2.4.3) implies that there exists a timestep $t^* \in [T]$ so that the iterate $\mu^{(t^*)}$ converges to the NE.

In the following, we introduce the key concept of regret decomposition, the (Local Regret) in infoset $s \in \mathcal{S}_i$ of player $i \in [2]$. Before introducing (Local Regret), we will define the counterfactual values for the regularized game (2.5.4). For any player $i \in [2]$ and infoset $s \in \mathcal{S}_i$, we have

$$\mathrm{CF}_i^{\psi^\Pi,\boldsymbol{\pi}}(s,a) \coloneqq \mathrm{CF}_i^{\boldsymbol{\pi}}(s,a) - \tau \sum_{s' \in \mathcal{S}_i \,:\, (s,a) \sqsubseteq s'} \frac{\mu_i^{\pi_i}(\sigma(s'))}{\mu_i^{\pi_i}(s,a)} \psi_{s'}^{\Delta}(\pi_i(\cdot \,|\, s')). \tag{4.0.1}$$

Then, we can define the local difference and regret for each infoset $s \in \mathcal{S}_i$ as,

$$G_s^{\psi^\Pi,(T)}(\pi_i) \coloneqq \sum_{t=1}^{T} \left(\left\langle \mathrm{CF}_i^{\psi^\Pi,\boldsymbol{\pi}^{(t)}}(s,\cdot), \pi_i(\cdot \,|\, s) - \pi_i^{(t)}(\cdot \,|\, s) \right\rangle + \tau \psi_s^{\Delta}(\pi_i^{(t)}(\cdot \,|\, s)) - \tau \psi_s^{\Delta}(\pi_i(\cdot \,|\, s)) \right) \tag{Local Difference}$$

$$R_s^{\psi^\Pi,(T)} \coloneqq \max_{\widehat{\pi}_i(\cdot \,|\, s) \in \Delta^{\mathcal{A}_s}} G_s^{\psi^\Pi,(T)}(\widehat{\pi}_i). \tag{Local Regret}$$

Here is the generalized version of the regret decomposition lemma [Farina et al., 2019b] for the regularized game.

This chapter builds on Section 5 of Liu et al. [2023].

Lemma 4.0.1. For any sequence of sequence-form strategy profiles $\mu^{(1)}, \mu^{(2)}, ..., \mu^{(T)}$ with length $T > 0$ and sequence-form strategy $\mu_i^{\pi_i} \in \Pi_i$ of player $i \in [2]$, we have

$$G_i^{\psi^\Pi,(T)}(\mu_i^{\pi_i}) = \sum_{s \in \mathcal{S}_i} \mu_i^{\pi_i}(\sigma(s)) G_s^{\psi^\Pi,(T)}(\pi_i) \tag{4.0.2}$$

$$R_i^{\psi^\Pi,(T)} \leq \max_{\mu_i^{\widehat{\pi}_i} \in \Pi_i} \sum_{s \in \mathcal{S}_i} \mu_i^{\widehat{\pi}_i}(\sigma(s)) R_s^{\psi^\Pi,(T)}. \tag{4.0.3}$$

The proof is postponed to Appendix B. Lemma 4.0.1 implies that when $R_s^{\psi^\Pi,(T)}$ is sublinear for any infoset $s \in \mathcal{S}$, the global regret $R_i^{\psi^\Pi,(T)}$ will be sublinear. Moreover, $R_s^{\psi^\Pi,(T)}$ is exactly the external regret in the online learning literature [Hazan et al., 2016], which is well-studied and there are many algorithms to solve it. Next, we will show the main result: sublinear $R_i^{\psi^\Pi,(T)}$ implies best-iterate convergence to the NE $\mu^{\tau,*}$ of the regularized game (2.5.4).

Theorem 4.0.2. For any sequence of strategy profiles $\mu^{(1)}, \mu^{(2)}, ..., \mu^{(T)}$ with length $T > 0$, we have

$$\sum_{t=1}^{T} D_{\psi^\Pi}\left(\mu^{(t)}, \mu^{\tau,*}\right) \leq \frac{R_1^{\psi^\Pi,(T)} + R_2^{\psi^\Pi,(T)}}{\tau}. \tag{4.0.4}$$

By pigeon-hole principle, there exists $t^* \in [T]$ such that

$$D_{\psi^\Pi}\left(\mu^{(t^*)}, \mu^{\tau,*}\right) \leq \frac{R_1^{\psi^\Pi,(T)} + R_2^{\psi^\Pi,(T)}}{\tau T}. \tag{4.0.5}$$

The proof of Theorem 4.0.2 follows from Lemma 3.0.5. By definition of $R_i^{\psi^\Pi,(T)}$, we have

$$\begin{aligned} R_i^{\psi^\Pi,(T)} + R_i^{\psi^\Pi,(T)} \geq & \sum_{t=1}^{T} \left(\left\langle F(\mu^{(t)}), \mu^{(t)} - \mu^{\tau,*} \right\rangle + \tau \psi^\Pi(\mu^{(t)}) - \tau \psi^\Pi\left(\mu^{\tau,*}\right) \right) \\ \overset{(i)}{\geq} & \tau \sum_{t=1}^{T} D_{\psi^\Pi}\left(\mu^{(t)}, \mu^{\tau,*}\right). \end{aligned}$$

(i) is by Lemma 3.0.5 and the proof is concluded. □

Theorem 4.0.2 implies that we can add additional regularization to any CFR variants, such as Zinkevich et al. [2007], Tammelin et al. [2015], Brown and Sandholm [2019a], and Farina et al. [2019a], to enable them best-iterate convergence. By choosing α_s in the dilated regularizer ψ^Π carefully [Lee et al., 2021], we can ensure ψ^Π to be 1-strongly convex and thus $D_{\psi^\Pi}\left(\mu^{(t)}, \mu^{\tau,*}\right) \geq \left\|\mu^{(t)} - \mu^{\tau,*}\right\|^2$.

Chapter 5

Q-Function based Regret Minimization (QFR)

In this chapter, we propose our policy gradient algorithm for EFGs, which we coin *Q-Function based Regret Minimization (*QFR*)*. In QFR, for each player $i \in [2]$ and state $s \in \mathcal{S}_i$, we enforce the strategy $\pi_i^{(t)}(\cdot \mid s)$ to explore with probability γ_s using the exploration strategy $\boldsymbol{\nu}_s$, in order to ensure that each infoset will be reached with a positive probability $\gamma > 0$.

Then, we show that QFR converges in best iterate to the *regularized Nash equilibrium*. Specifically, QFR will converge to the solution $\boldsymbol{\mu}^{(\tau,\gamma),*} = (\mu_1^{(\tau,\gamma),*}, \mu_2^{(\tau,\gamma),*})$ of the original bilinear minimax objective plus additional regularization term $\psi_{\mathrm{bi}}^{\Pi_1}(\mu_1^{\pi_1}, \mu_2^{\pi_2})$ and $\psi_{\mathrm{bi}}^{\Pi_2}(\mu_1^{\pi_1}, \mu_2^{\pi_2})$. $\psi_{\mathrm{bi}}^{\Pi_1}(\mu_1^{\pi_1}, \mu_2^{\pi_2})$ is strongly convex with respect to $\mu_1^{\pi_1}$ and convex with respect to $\mu_2^{\pi_2}$. Conversely, $\psi_{\mathrm{bi}}^{\Pi_2}(\mu_1^{\pi_1}, \mu_2^{\pi_2})$ is strongly convex with respect to $\mu_2^{\pi_2}$ and convex with respect to $\mu_1^{\pi_1}$. In contrast to the original bilinear objective, optimizing the regularized objective will stabilize the training process, and result in better convergence results. Formally, the *regularized and perturbed* (perturb refers to the exploration) game is,

$$\min_{\substack{\mu_1^{\pi_1} \in \Pi_1 : \\ \forall s \in \mathcal{S}_1,\ \pi_1(\cdot \mid s) \in \Delta_{\gamma_s, \boldsymbol{\nu}_s}^{|\mathcal{A}_s|}}} \ \max_{\substack{\mu_2^{\pi_2} \in \Pi_2 : \\ \forall s \in \mathcal{S}_2,\ \pi_2(\cdot \mid s) \in \Delta_{\gamma_s, \boldsymbol{\nu}_s}^{|\mathcal{A}_s|}}} \left(\mu_1^{\pi_1}\right)^\top \boldsymbol{A} \mu_2^{\pi_2} + \tau \psi_{\mathrm{bi}}^{\Pi_1}(\mu_1^{\pi_1}, \mu_2^{\pi_2}) - \tau \psi_{\mathrm{bi}}^{\Pi_2}(\mu_1^{\pi_1}, \mu_2^{\pi_2}) \quad (5.0.1)$$

where $\Delta_{\gamma_s, \boldsymbol{\nu}_s}^{|\mathcal{A}_s|} \coloneqq \left\{ \boldsymbol{u} \in \Delta^{|\mathcal{A}_s|} : \forall a \in \mathcal{A}_s, \quad u_a \geq \gamma_s \nu_{s,a} \right\}$, and $\tau \geq 0$ controls the magnitude of the regularizer. Note that the NE of the non-regularized game can be computed by annealing the regularization coefficient τ as Section 3.1.

To decompose the regularizer $\psi_{\mathrm{bi}}^{\Pi_i}$ to each infoset for efficient update, we resort to the concept *dilated regularizer* [Hoda et al., 2010]. Take Euclidean norm for example, unlike naively choosing $\frac{1}{2} \left\| \mu_i^{\pi_i} \right\|^2$ as the regularizer, dilated regularizer weights the regularizer $\frac{1}{2} \left\| \pi_i(\cdot \mid s) \right\|^2$ at each infoset $s \in \mathcal{S}_i$ by the reach probability of player i to s, *i.e.*, $\frac{1}{2} \sum_{s \in \mathcal{S}_i} \mu_i^{\pi_i}(\sigma(s)) \left\| \pi_i(\cdot \mid s) \right\|^2$. It is shown in Hoda et al. [2010] that the dilated regularizer is strongly convex with respect to $\mu_i^{\pi_i}$. However, dilated regularizer ψ^{Π_i} only weights the reach probability of player i, neglecting that of the opponents, of which the asymmetry causes importance sampling when estimating the regularizer term in a rolling trajectory. Therefore, a natural solution is weighting the strongly convex regularizer $\psi_s^\Delta \colon \Delta^{|\mathcal{A}_s|} \to \mathbb{R}$ at each infoset $s \in \mathcal{S}$ (typically it is Euclidean norm or negative entropy) by the reach probability of all players.

This chapter builds on Liu et al. [2024b].

The details can be referred to Appendix C.1. With all ingredients at hand, we can introduce the algorithm `Q-Function based Regret minimization (QFR)`. At each timestep t, `QFR` will sample a trajectory according to the current strategy $\boldsymbol{\pi}^{(t-1)}$ (Line 3 of Algorithm 1). Then, the *trajectory Q-value* will be estimated from the trajectory (Line 10-12, ψ estimates the trajectory Q-value contributed by additional bidilated regularizer of the game, W estimates the that contributed by the reward of the original game). Lastly, all infosets along the trajectory will be updated with a variant of Regularized Optimistic Mirror Descent (Reg-OMD) proposed in Liu et al. [2023] (Line 14). In Reg-OMD, at timestep t, the strategy $\boldsymbol{\pi}^{(t)}$ serves as the prediction of $\overline{\boldsymbol{\pi}}^{(t+1)}$. Then in the next timestep $t+1$, $\overline{\boldsymbol{\pi}}^{(t)}$ will be updated with the trajectory Q-value estimated at $\boldsymbol{\pi}^{(t)}$. The pseudocode of `QFR` is proposed in Algorithm 1.

Algorithm 1 Q-Function based Regret minimization (QFR)

1: Initialize $\pi_i^{(1)}(\cdot \mid s), \overline{\pi}_i^{(1)}(\cdot \mid s)$ as uniform distribution over $\Delta^{|\mathcal{A}_s|}$ for any $i \in [2]$ and $s \in \mathcal{S}_i$.
2: **for** $t = 2, 3, \cdots, T$ **do**
3: Sample a trajectory $(h_0, a_0, h_1, \cdots, h_K) \sim \boldsymbol{\pi}^{(t-1)}$, where $h_0 = \emptyset$ and h_K is terminal
4: Let $s_0, s_1, \cdots, s_K$ be the infosets corresponding to $h_0, h_1, \cdots, h_K$
5: **for** $k = K, K-1, \cdots, 0$ **do**
6: **if** node h_k belongs to the chance player **then**
7: **continue** to the new iteration directly
8: **else**
9: $p \leftarrow p(s_k)$
10: $\psi \leftarrow -\sum_{k'=k+1:p(s_{k'})=p}^{K} \psi_{s_{k'}}^{\Delta}(\pi_p^{(t-1)}) + \sum_{k'=k+1:p(s_{k'})=3-p}^{K} \psi_{s_{k'}}^{\Delta}(\pi_{3-p}^{(t-1)})$
11: $W \leftarrow \mathcal{U}_p(h_K)$, end-of-trajectory utility of player p (only terminal nodes' utility $\neq 0$)
12: Compute the unbiased estimator of trajectory Q-value as

$$\widetilde{q}^{(t-1)}(s_k, a) = \begin{cases} (W + \tau\psi)\, \nabla_a \log \pi_p^{(t-1)}(a \mid s_k) & a = a_k \\ 0 & \text{Otherwise.} \end{cases}$$

13: Compute the estimated value function as

$$\widetilde{V}^{\pi_p}(s_k) = \mathbb{E}_{a \sim \pi_p(\cdot \mid s_k)}\left[\widetilde{q}^{(t-1)}(s_k, a)\right] - \tau \psi_{s_k}^{\Delta}\left(\pi_p(\cdot \mid s_k)\right).$$

14: Update $\overline{\pi}_p^{(t-1)}, \pi_p^{(t-1)}$ according to

$$\begin{aligned} \overline{\pi}_p^{(t)}(\cdot \mid s_k) &\leftarrow \operatorname*{argmax}_{\pi_p(\cdot \mid s_k) \in \Delta_{\gamma_{s_k}, \boldsymbol{\nu}_{s_k}}^{|\mathcal{A}_{s_k}|}} \widetilde{V}^{\pi_p}(s_k) - \frac{1}{\eta_{s_k}} D_{\psi_{s_k}^{\Delta}}\left(\pi_p(\cdot \mid s_k), \overline{\pi}_p^{(t-1)}(\cdot \mid s_k)\right) \\ \pi_p^{(t)}(\cdot \mid s_k) &\leftarrow \operatorname*{argmax}_{\pi_p(\cdot \mid s_k) \in \Delta_{\gamma_{s_k}, \boldsymbol{\nu}_{s_k}}^{|\mathcal{A}_{s_k}|}} \widetilde{V}^{\pi_p}(s_k) - \frac{1}{\eta_{s_k}} D_{\psi_{s_k}^{\Delta}}\left(\pi_p(\cdot \mid s_k), \overline{\pi}_p^{(t)}(\cdot \mid s_k)\right), \end{aligned} \tag{5.0.2}$$

where $\Delta_{\gamma_s, \boldsymbol{\nu}_s}^{|\mathcal{A}_s|} := \left\{\boldsymbol{u} \in \Delta^{|\mathcal{A}_s|} : \forall a \in \mathcal{A}_s, \quad u_a \geq \gamma_s \nu_{s,a}\right\}$

15: For all infosets s not visited at timestep t, let $\pi_{p(s)}^{(t)}(\cdot \mid s) = \pi_{p(s)}^{(t-1)}$ (same for $\overline{\pi}_{p(s)}^{(t)}$)

Remark 5.0.1. One may argue that there is still importance sampling in Algorithm 1 because $\nabla_a \log \pi_p^{(t-1)}(a \mid s_k) = \frac{1}{\pi_p^{(t-1)}(a \mid s_k)}$. However, in contrast to previous work in which

the importance sampling factor, $\frac{1}{\mu_p^{(t-1)}(s_k,a)}$, can be as large as the game size, here it is simply proportional to the action set size. Moreover, such importance sampling can be easily side-stepped by sampling additional trajectories starting from every infoset-action pair in $\{(s_k,a)\}_{k\in\{0,1,\dots,K\},a\in\mathcal{A}_s}$ to compute each component of $\widetilde{q}^{(t-1)}(s_k,a)$. The price is sampling $\mathcal{O}(\mathcal{D}^2\max_{s\in\mathcal{S}}|\mathcal{A}_s|)$ nodes in each iteration instead of $\mathcal{O}(\mathcal{D})$ in Algorithm 1.

We define the largest learning rate among all ancestor infosets of $s\in\mathcal{S}$ as $\eta_s^{\text{anc}} := \max_{i\in[2],h\in s}\max_{(s',a')\sqsubseteq\sigma_i(h)}\eta_{s'}$ (η_s is the learning rate of infoset s), and we have the following theorem that establishes the best-iterate convergence of Algorithm 1 to the NE $\boldsymbol{\mu}^{(\tau,\gamma),*} = (\mu_1^{(\tau,\gamma),*},\mu_2^{(\tau,\gamma),*})$ of (5.0.1) with high-probability.

Theorem 5.0.2 (Informal)**.** Consider Algorithm 1. When $\frac{\eta_s^{\text{anc}}}{\eta_s}\le\tau C_s^{\eta,T}$, where C_s^{η} is a game-dependent constant, and η_s is smaller than a game dependent constant (formally defined in Appendix C.4) for any $s\in\mathcal{S}$, we have the following guarantee with probability $1-2\delta$.

$$\sum_{t=2}^{T} D_{\psi^{\Pi}}(\boldsymbol{\mu}^{(\tau,\gamma),*},\boldsymbol{\mu}^{\overline{\boldsymbol{\pi}}^{(t)}}) \le \widetilde{\mathcal{O}}\left(\max_{s\in\mathcal{S}}\eta_s T\right) + \widetilde{\mathcal{O}}\left(\frac{1}{\min_{s\in\mathcal{S}}\eta_s}\right) + \widetilde{\mathcal{O}}\left(\sqrt{T\log\frac{1}{\delta}}\right). \tag{5.0.3}$$

The $\widetilde{\mathcal{O}}$ notion hides the logarithm of T. The proof and the formal version are postponed to Appendix C.4. Theorem 5.0.2 gives a high-probability upper-bound on the cumulative Bregman divergence. By letting $\eta_s=\Theta(1/\sqrt{T})$ for all infoset $s\in\mathcal{S}$, the right-hand-side of (5.0.3) is bounded by $\widetilde{\mathcal{O}}(\sqrt{T})$. Therefore, it implies that $\sum_{t=2}^{T} D_{\psi^{\Pi}}(\boldsymbol{\mu}^{(\tau,\gamma),*},\boldsymbol{\mu}^{\overline{\boldsymbol{\pi}}^{(t)}})$ is upper-bounded by $\widetilde{\mathcal{O}}(\sqrt{T})$ with probability $1-2\delta$. Then, there must exist some $t^*\in[T]$ so that $D_{\psi^{\Pi}}(\boldsymbol{\mu}^{(\tau,\gamma),*},\boldsymbol{\mu}^{\overline{\boldsymbol{\pi}}^{(t^*)}})\le\widetilde{\mathcal{O}}(1/\sqrt{T})$, because the minimum over $\left\{D_{\psi^{\Pi}}(\boldsymbol{\mu}^{(\tau,\gamma),*},\boldsymbol{\mu}^{\overline{\boldsymbol{\pi}}^{(t)}})\colon t=2,3,\cdots,T\right\}$ must be bounded by the average, which is $\widetilde{\mathcal{O}}(1/\sqrt{T})$. Therefore, by computing the exploitability (the expected utility confronting a best-responding opponent) routinely, we can find an approximate NE [Liu et al., 2023] of the regularized game. Moreover, according to Lemma Liu et al. [2023, Lemma D.1.], the exploitability of the regularized NE will be bounded by $\mathcal{O}(\tau)$ in the original game so that our iterates will also get a low exploitability in the original game by fixing τ to be small or anneal it as Liu et al. [2023].

5.1 Analysis

In this section, we provide the proof sketch of Theorem 5.0.2. Section 5.1.1 introduces some necessary notions and properties for the analysis. Section 5.1.2 shows the convergence of `QFR` under full-information feedback (traversing all infosets at each iteration), and Section 5.1.3 generalizes to the stochastic setting discussed above.

5.1.1 Preliminaries and Basic Properties

In order to keep the presentation modular between Q-values and trajectory Q-values, we will assume that, at each iteration t, the local strategy at each infoset will be updated by taking a step in the direction of some generalized *value vector* $q^{(t)}(s,\cdot)\in\mathbb{R}^{|\mathcal{A}_s|}$. For each $a\in\mathcal{A}_s$, (we

will use $\mathsf{CF}_1^{(t)}(s,a)$ instead of $\mathsf{CF}_1^{\boldsymbol{\pi}^{(t)}}(s,a)$ and $\mu_i^{(t)}$ instead of $\mu_i^{\pi_i^{(t)}}$ as the shorthand notion), the relationship of counterfactual values and the feedback is,

$$\mathsf{CF}_1^{(t)}(s,a) = \begin{cases} \sum_{h\in s} \mu_c(h)\mu_2^{(t)}(\sigma_2(h)) \cdot q^{(t)}(s,a) & q^{(t)}(s,\cdot) \text{ is Q-value} \\ \frac{q^{(t)}(s,a)}{\mu_1^{(t)}(\sigma(s))} & q^{(t)}(s,\cdot) \text{ is trajectory Q-value} \\ q^{(t)}(s,a) & q^{(t)}(s,\cdot) \text{ is counterfactual value} \end{cases} \tag{5.1.1}$$

It is noteworthy that when sampling a trajectory from the root to a terminal node, the utility will be a good estimator of trajectory Q-value. Therefore, to estimate $\mathsf{CF}_1^{(t)}(s,a)$, we need to divide the reaching probability $\mu_1^{(t)}(\sigma(s))$ of s, which can be extremely small and thus induces a large variance. In the following, we will write $\mathsf{CF}_1^{(t)}(s,a) = m_s^{(t)} q^{(t)}(s,a)$ and $m_s^{(t)}$ is different for different types of $q^{(t)}(s,\cdot)$ according to (5.1.1).

Note that the value of $m_s^{(t)}$ in most cases depends on the strategies that are produced by the algorithm; hence, there is some circularity in the dependence between the properties satisfied by $m_s^{(t)}$ and those satisfied by our algorithm. To break this circularity, at the heart of our correctness proof we will verify and leverage two key properties of the sequence of $m_s^{(t)}$ that arises from using our algorithm: *boundedness* and *stability*, as detailed next.

Property 1 (Boundedness)**.** For any $t \in [T]$ and $s \in \mathcal{S}$, we have $m_s^{(t)} \in [M_1, M_2]$ where $0 < M_1 \le M_2 < +\infty$.

Property 2 (Stability)**.** For any $t \in [T-1]$ and $s \in \mathcal{S}$, we define the largest learning rate among all ancestor infosets of s as $\eta_s^{\text{anc}} := \max_{i\in[2],h\in s} \max_{(s',a')\sqsubseteq\sigma_i(h)} \eta_{s'}$, where $\eta_{s'}$ is the learning rate of infoset s', then

$$\left|m_s^{(t+1)} - m_s^{(t)}\right| \le C_s^- \eta_s^{\text{anc}} \qquad \text{(Additive Stability)}$$

$$\left|\frac{m_s^{(t+1)}}{m_s^{(t)}} - 1\right| \le C_s' \eta_s^{\text{anc}}. \qquad \text{(Multiplicative Stability)}$$

Property 1 will be satisfied by enforcing $\pi^{(t)}(\cdot \mid s) \succeq \gamma_s \boldsymbol{\nu}_s$ for every $s \in \mathcal{S}$, where $\gamma_s \in (0,1]$ and $\boldsymbol{\nu}_s \in \Delta^{|\mathcal{A}_s|}$ are specified in Appendix C.5.2. The proof that our algorithm produces iterates that satisfy Property 2 can be found in Appendix C.2.

5.1.2 Convergence with Full Information Feedback

`QFR` runs a variant of Regularized Optimistic Mirror Descent (Reg-OMD) [Liu et al., 2023] algorithm to update the strategy in each infoset. For notational simplicity, we define the reach probability of opponents to an infoset $s \in \mathcal{S}$ as $\mu_{-p(s)}^{(t)}(s) := \sum_{h\in s} \mu_c(h)\mu_{3-p(s)}^{(t)}(\sigma_{3-p(s)}(h))$.

The update rule is

$$\begin{aligned}\pi_{p(s)}^{(t)}(\cdot\,|\,s) = &\operatorname*{argmin}_{\pi_{p(s)}(\cdot\,|\,s)\in\Delta^{|\mathcal{A}_s|}_{\gamma_s,\boldsymbol{\nu}_s}} \left\langle \pi_{p(s)}(\cdot\,|\,s), -q^{(t-1)}(s,\cdot)\right\rangle + \frac{\tau\mu^{(t-1)}_{-p(s)}(s)}{m_s^{(t-1)}}\psi_s^{\Delta}(\pi_{p(s)}(\cdot\,|\,s)) \\ &\qquad + \frac{1}{\eta_s}D_{\psi_s^{\Delta}}(\pi_{p(s)}(\cdot\,|\,s), \bar{\pi}^{(t)}_{p(s)}(\cdot\,|\,s)) \\ \bar{\pi}_{p(s)}^{(t+1)}(\cdot\,|\,s) = &\operatorname*{argmin}_{\pi_{p(s)}(\cdot\,|\,s)\in\Delta^{|\mathcal{A}_s|}_{\gamma_s,\boldsymbol{\nu}_s}} \left\langle \pi_{p(s)}(\cdot\,|\,s), -q^{(t)}(s,\cdot)\right\rangle + \frac{\tau\mu^{(t)}_{-p(s)}(s)}{m_s^{(t)}}\psi_s^{\Delta}(\pi_{p(s)}(\cdot\,|\,s)) \\ &\qquad + \frac{1}{\eta_s}D_{\psi_s^{\Delta}}(\pi_{p(s)}(\cdot\,|\,s), \bar{\pi}^{(t)}_{p(s)}(\cdot\,|\,s))\end{aligned} \tag{5.1.2}$$

where $\Delta^{|\mathcal{A}_s|}_{\gamma_s,\boldsymbol{\nu}_s} := \left\{\boldsymbol{u}\in\Delta^{|\mathcal{A}_s|}\colon \forall a\in\mathcal{A}_s,\ \ u_a\geq\gamma_s\nu_{s,a}\right\}$ and ψ_s^{Δ} is the regularizer chosen for infoset s. Here $q^{(t)}(s,\cdot)$ can be the trajectory Q-value, Q-value, or counterfactual value associated with $\boldsymbol{\pi}^{(t)}$. When $q^{(t)}(s,\cdot)$ is the trajectory Q-value, (5.1.2) is the full-information version of (5.0.2).

By analyzing the update rule (5.1.2), we can get the following inequality. For any $s\in\mathcal{S}$ and strategy $\pi_{p(s)}(\cdot\,|\,s)\in\Delta^{|\mathcal{A}_s|}_{\gamma_s,\boldsymbol{\nu}_s}$, we have

$$\begin{aligned}&\sum_{t=1}^{T}\left(\tau\mu^{(t)}_{-p(s)}(s)\psi_s^{\Delta}(\pi^{(t)}_{p(s)}(\cdot\,|\,s)) - \tau\mu^{(t)}_{-p(s)}(s)\psi_s^{\Delta}(\pi_{p(s)}(\cdot\,|\,s)) + m_s^{(t)}\langle -q^{(t)}(s,\cdot), \pi^{(t)}_{p(s)}(\cdot\,|\,s) - \pi_{p(s)}(\cdot\,|\,s)\rangle\right) \\ \leq &\sum_{t=2}^{T}\underbrace{\left(\frac{m_s^{(t)}-m_s^{(t-1)}}{\eta_s} - \tau\mu^{(t-1)}_{-p(s)}(s)\right)}_{①} D_{\psi_s^{\Delta}}(\pi_{p(s)}(\cdot\,|\,s), \bar{\pi}^{(t)}_{p(s)}(\cdot\,|\,s)) + \mathcal{O}(\eta_s T) + \mathcal{O}\left(\frac{1}{\eta_s}\right).\end{aligned} \tag{5.1.3}$$

Then, since $\left|m_s^{(t)} - m_s^{(t-1)}\right| \leq \mathcal{O}(\eta_s^{\text{anc}})$, $① \leq -\frac{\tau}{2}\mu^{(t-1)}_{-p(s)}(s) \leq -\frac{\tau\gamma}{2}\sum_{h\in s}\mu_c(h)$ when η_s^{anc} is smaller then η_s by a small enough constant (please refer to Appendix C.3 for details). Then, by applying the generalized regret decomposition lemma [Liu et al., 2023] (details can be found in Lemma C.3.3) to (5.1.3), the difference of $\pi^{(t)}_{p(s)}$ and $\pi_{p(s)}$ in an infoset s can be extended to the difference of the whole game. Specifically, by letting

$$\begin{aligned}\text{diff}_1(\boldsymbol{\mu}^{\boldsymbol{\pi}}, \boldsymbol{\mu}^{\boldsymbol{\pi}'}) \coloneqq &\left(\mu_1^{\pi_1} - \mu_1^{\pi_1'}\right)^{\top}\boldsymbol{A}\mu_2^{\pi_2} - \tau\left(\psi_{\text{bi}}^{\Pi_1}(\mu_1^{\pi_1}, \mu_2^{\pi_2}) - \psi_{\text{bi}}^{\Pi_1}(\mu_1^{\pi_1'}, \mu_2^{\pi_2})\right) \\ &+ \tau\left(\psi_{\text{bi}}^{\Pi_2}(\mu_1^{\pi_1}, \mu_2^{\pi_2}) - \psi_{\text{bi}}^{\Pi_2}(\mu_1^{\pi_1'}, \mu_2^{\pi_2})\right)\end{aligned}$$

and diff_2 similarly, then the summation of the left-hand-side of (5.1.3) over all infoset $s\in\mathcal{S}$ is equal to is equal to

$$\sum_{t=1}^{T}\text{diff}_1(\boldsymbol{\mu}^{(\tau,\gamma),*}, \boldsymbol{\mu}^{(t)}) + \text{diff}_2(\boldsymbol{\mu}^{(\tau,\gamma),*}, \boldsymbol{\mu}^{(t)}) \geq 0. \tag{5.1.4}$$

The non-negativity is because $\boldsymbol{\mu}^{(\tau,\gamma),*}$ is the NE of (5.0.1). By combining (5.1.3) and (5.1.4), we have

$$0 \le \sum_{s\in\mathcal{S}} \mu^{(\tau,\gamma),*}(\sigma(s)) \left(-\frac{\tau\gamma}{2} \sum_{h\in s} \mu_c(h) \sum_{t=2}^{T} D_{\psi_s^\Delta}(\pi_{p(s)}^{(\tau,\gamma),*}(\cdot \mid s), \bar{\pi}_{p(s)}^{(t)}(\cdot \mid s)) + \mathcal{O}(\eta_s T) + \mathcal{O}\left(\frac{1}{\eta_s}\right) \right)$$
$$\overset{(a)}{=} -\frac{\tau\gamma}{2} \min_{s\in\mathcal{S}} \sum_{h\in s} \mu_c(h) \sum_{t=2}^{T} D_{\psi^\Pi}(\boldsymbol{\mu}^{(\tau,\gamma),*}, \boldsymbol{\mu}^{\bar{\boldsymbol{\pi}}^{(t)}}) + \mathcal{O}(\max_{s\in\mathcal{S}} \eta_s T) + \mathcal{O}\left(\frac{1}{\min_{s\in\mathcal{S}} \eta_s}\right).$$

(a) is by Lemma 5.1.1 in the following. By rearranging the terms, we can get an upperbound on $\sum_{t=2}^{T} D_{\psi^\Pi}(\boldsymbol{\mu}^{(\tau,\gamma),*}, \boldsymbol{\mu}^{\bar{\boldsymbol{\pi}}^{(t)}})$.

Lemma 5.1.1 (Lemma D.2. in Liu et al. [2022])**.** For any strategy $\boldsymbol{\mu}^{\boldsymbol{\pi}}, \boldsymbol{\mu}^{\widetilde{\boldsymbol{\pi}}} \in \Pi$ and regularizer $\psi_s^\Delta \colon \Delta^{|\mathcal{A}_s|} \to \mathbb{R}$ for each infoset $s \in \mathcal{S}$, we have

$$D_{\psi^\Pi}(\boldsymbol{\mu}^{\boldsymbol{\pi}}, \boldsymbol{\mu}^{\widetilde{\boldsymbol{\pi}}}) = \sum_{s\in\mathcal{S}} \mu^{\boldsymbol{\pi}}(\sigma(s)) D_{\psi_s^\Delta}(\pi_{p(s)}(\cdot \mid s), \widetilde{\pi}_{p(s)}(\cdot \mid s)). \tag{5.1.5}$$

For completeness, the proof of the lemma can be found in Appendix C.3.1. Here is the full theorem.

Theorem 5.1.2 (Informal)**.** Consider the update rule (5.1.2) and $q^{(t)}(s,\cdot)$ is chosen to be counterfactual value, trajectory Q-value, or Q-value. When $\frac{\eta_s^{\text{anc}}}{\eta_s} \le \tau C_s^\eta$, where C_s^η is a game-dependent constant, and η_s is smaller than a game dependent constant (formally defined in Appendix C.3) for any $s \in \mathcal{S}$, we have the following guarantee.

$$\sum_{t=2}^{T} D_{\psi^\Pi}(\boldsymbol{\mu}^{(\tau,\gamma),*}, \boldsymbol{\mu}^{\bar{\boldsymbol{\pi}}^{(t)}}) \le \mathcal{O}\left(\max_{s\in\mathcal{S}} \eta_s T\right) + \mathcal{O}\left(\frac{1}{\min_{s\in\mathcal{S}} \eta_s}\right). \tag{5.1.6}$$

The proof and the formal version are postponed to Appendix C.3. Therefore, by choosing $\eta_s = \Theta\left(\frac{1}{\sqrt{T}}\right)$ for any $s \in \mathcal{S}$ as in Theorem 5.1.2, `QFR` enjoys best-iterate convergence with full-information feedback.

5.1.3 Convergence with Stochastic Feedback

We complement the results of Section 5.1.2 by showing that the best-iterate convergence guaranteed by `QFR` is still guaranteed when only visiting a trajectory at each iteration. The proof utilizes standard concentration inequalities, incurring an additional sublinear cost caused by the noise incurred from sampling, as recalled in the following lemma.

Lemma 5.1.3 (Generalization of Proposition 1 in Farina et al. [2020])**.** Let $M, \widetilde{M}$ be positive constants such that $\left|f^{(t)}(\boldsymbol{u}) - f^{(t)}(\boldsymbol{u}')\right| \le M$ and $\left|\widetilde{f}^{(t)}(\boldsymbol{u}) - \widetilde{f}^{(t)}(\boldsymbol{u}')\right| \le \widetilde{M}$ for any $\boldsymbol{u}, \boldsymbol{u}' \in \mathcal{C}$ for any $t \in [T]$, where $\mathcal{C}$ is a convex set. Then, if for any $\boldsymbol{u}$, $\mathbb{E}[\widetilde{f}^{(t)}(\boldsymbol{u}) \mid \widetilde{f}^{(1)}, \widetilde{f}^{(2)}, \cdots, \widetilde{f}^{(t-1)}] = f^{(t)}(\boldsymbol{u})$ and $\boldsymbol{u}^{(t)}$ is deterministically influenced by $\widetilde{f}^{(1)}, \widetilde{f}^{(2)}, \cdots, \widetilde{f}^{(t-1)}$, then for any $\delta \in (0,1)$ and $\boldsymbol{u} \in \mathcal{C}$, we have

$$\Pr\left(\sum_{t=1}^{T} \left(f^{(t)}(\boldsymbol{u}) - f^{(t)}(\boldsymbol{u}^{(t)})\right) \le \sum_{t=1}^{T} \left(\widetilde{f}^{(t)}(\boldsymbol{u}) - \widetilde{f}^{(t)}(\boldsymbol{u}^{(t)})\right) + (M + \widetilde{M})\sqrt{2T \log \frac{1}{\delta}} \right) \ge 1 - \delta.$$

Next, we will substitute the following values into Lemma 5.1.3,

$$f_s^{(t)}(\boldsymbol{u}) \coloneqq \frac{1}{\mu_{p(s)}^{(t)}(\sigma(s))}\langle q^{(t)}(s,\cdot), \boldsymbol{u}\rangle - \tau\mu_{-p(s)}^{(t)}(s)\psi_s^{\Delta}(\boldsymbol{u})$$

$$\widetilde{f}_s^{(t)}(\boldsymbol{u}) \coloneqq \begin{cases} \frac{1}{\mu_{p(s)}^{(t)}(\sigma(s))}\langle \widetilde{q}^{(t)}(s,\cdot), \boldsymbol{u}\rangle - \frac{\tau}{\mu_{p(s)}^{(t)}(\sigma(s))}\psi_s^{\Delta}(\boldsymbol{u}) & s \text{ is visited at timestep } t \\ 0 & \text{Otherwise} \end{cases}$$

, where $\widetilde{q}^{(t)}$ is defined in Algorithm 1. The proof of $\widetilde{f}_s^{(t)}$ being an unbiased estimator of $f_s^{(t)}$ is postponed to Appendix C.4. Then, (5.1.3) can be bounded by $\sum_{t=1}^{T}\left(\widetilde{f}_s^{(t)}(\pi_{p(s)}(\cdot\,|\,s)) - \widetilde{f}_s^{(t)}(\pi_{p(s)}^{(t)}(\cdot\,|\,s))\right)$, which can be further bounded by analyzing the update-rule (5.0.2). The analysis is similar to the one in Section 5.1.2 and can be found in Appendix C.4. Finally, we have Theorem 5.0.2.

5.2 Experiments

In the following, we will introduce the experiment setup, the result under trajectory feedback (Algorithm 1), the result under full-information feedback, and the result when applying deep learning to approximate the strategy.

5.2.1 Experiment Setup

Figure 5.1 and Figure 5.2 are conducted on 240 cores of Intel Xeon Platinum 8260 and Figure 5.3 is conducted on Intel(R) Xeon Gold 6248 with NVidia Volta V100. The code is based on LiteEFG [Liu et al., 2024a] with game environments implemented by OpenSpiel [Lanctot et al., 2019].

In the experiments, we apply `QFR` in 4-Sided Liar's Dice, Leduc Poker [Southey et al., 2005], Kuhn Poker [Kuhn, 1950], and 2×2 Abrupt Dark Hex. The learning rate is the *same* in all infosets, unlike what the theorem requires, which shows that `QFR` is easier to implement than what the theory suggests. Note that for MMD [Sokota et al., 2023], there is no theory for convergence when using trajectory Q-value and Q-value as feedback, while `QFR` has.

In order to pick hyperparameters, we performed a grid-search for `QFR` and `MMD` on learning rate η, regularization τ, perturbation γ, and the regularizer is either negative entropy or Euclidean distance. For BalancedOMD (BOMD) [Bai et al., 2022] and BalancedFTRL (BFTRL) [Fiegel et al., 2023], we applied grid search to the learning rate η and fixed the exploration rate (IX parameter) to $\frac{\eta}{20}$ as suggested in Fiegel et al. [2023]. For the outcome-sampling CFR / CFR+, we also applies grid-search on the exploration parameter.

5.2.2 Experimental results under Trajectory Feedback

The experimental result of Algorithm 1 is presented in Figure 5.1. In the experiments, `QFR` and MMD are both using an unbiased estimator of the trajectory Q-value, while CFR and CFR+ use an unbiased estimator of the counterfactual value.

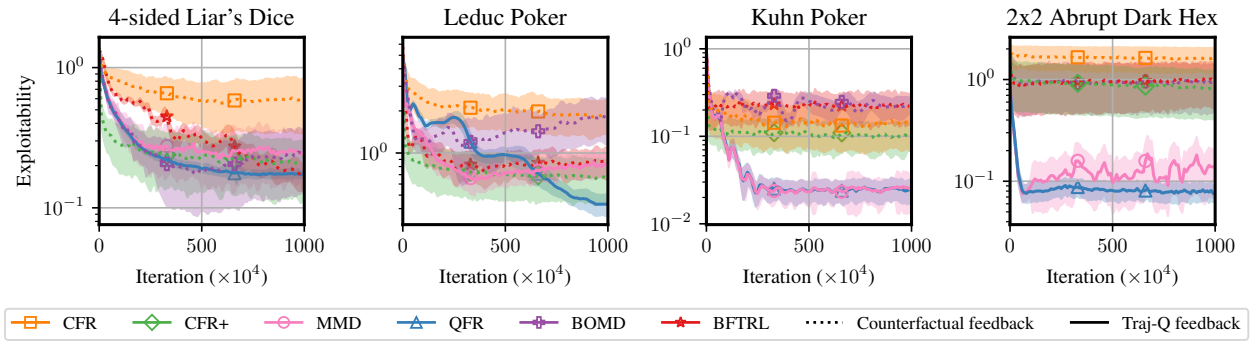

Figure 5.1: Exploitability of Algorithm 1 in 4 benchmark games. We can see that `QFR` outperforms outcome-sampling CFR / CFR+, MMD, and BOMD in all games. It outperforms BFTRL in all games except Liar's Dice. For each line, we repeat the experiments 100 times with different seeds.

Figure 5.1 shows that `QFR` outperforms outcome-sampling CFR, CFR+, and BOMD in all games. Moreover, `QFR` outperforms BFTRL in all games except Liar's Dice, with a relatively small gap (QFR 0.174 v.s. BFTRL 0.167 in exploitability). The reason may be Liar's Dice is too easy since it can be solved within 50 iterations with full-information feedback (see Figure 5.2). Lastly, `QFR` outperforms MMD in all games.

The superiority of `QFR` over CFR, CFR+, BOMD, and BFTRL may be attributed to both the additional regularization (best-iterate convergence) and the avoidance of importance sampling. For MMD, `QFR` is superior due to the optimistic updates, since optimistic updates allow predictions of the gradients at the next iteration.

5.2.3 Experimental results under Full-Information Feedback

We present the experimental results for full information feedback in Figure 5.2, that is, we use trajectory Q-value, Q-value, or counterfactual value as $q^{(t)}(s, \cdot)$ in the update rule (5.1.2). In Figure 5.2, we plot the last-iterate performance for all the algorithms to ensure the comparison is fair.

5.2.4 Experimental results for Deep Learning

In Figure 5.3, we present the ablation study on importance sampling when the strategy is approximated by the neural network. Our implementation of `QFR` is based on PPO [Schulman et al., 2017] in CleanRL [Huang et al., 2022]. We can see that with importance sampling, the network gradient blows up so that the network does not converge, even though we have applied gradient clipping.

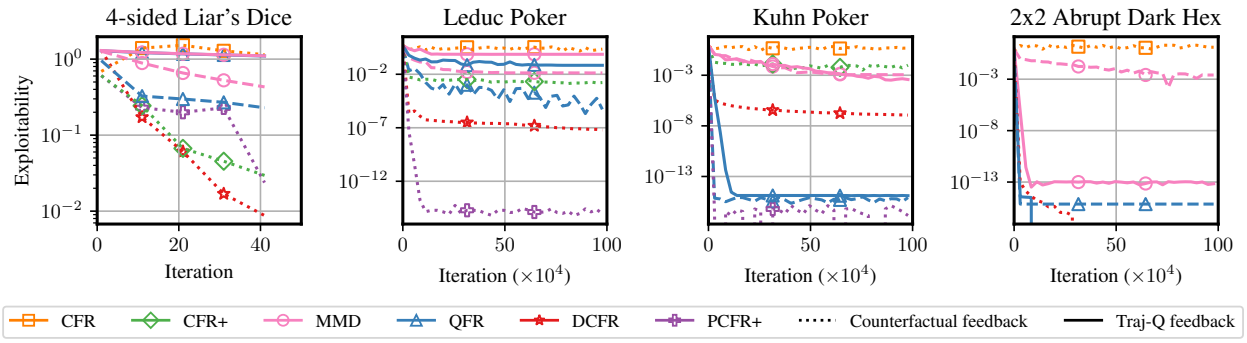

Figure 5.2: The result of full-information feedback in four benchmark games. We compare with CFR [Zinkevich et al., 2007], CFR+ [Tammelin et al., 2015], MMD [Sokota et al., 2023], DCFR [Brown and Sandholm, 2019a], and PCFR+ [Farina et al., 2021a]. We can see that QFR outperforms MMD in all games. However, due to multiplicative noise caused by using Q-values, QFR cannot outperform PCFR+, an advanced variant of CFR.

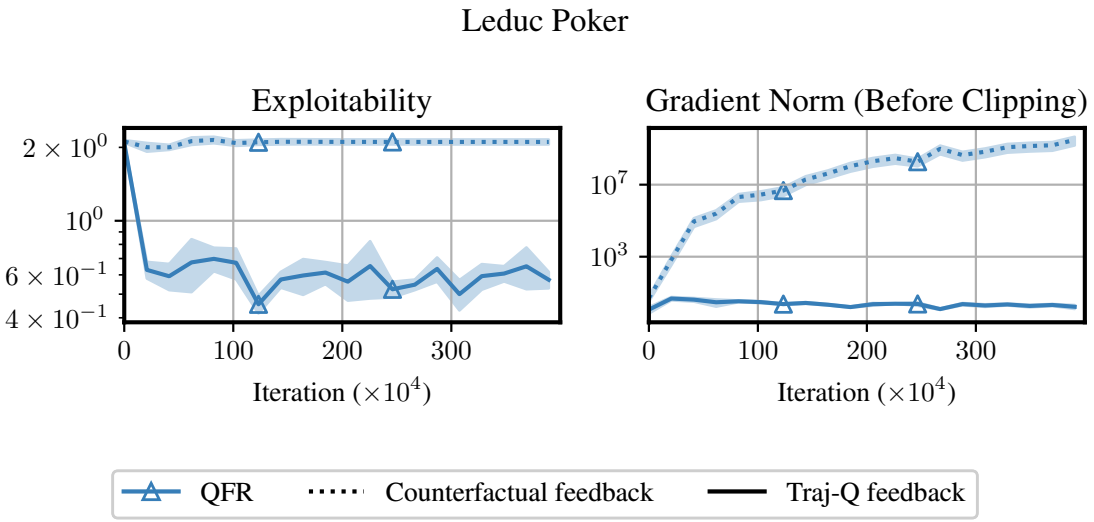

Figure 5.3: The result of `QFR` with sampling feedback. We can see that with importance sampling, the gradient norm keeps growing so that the network does not converge even with gradient clipping. The right figure shows the gradient **before clipping** and the gradient will be clipped so that its norm is bounded by 0.5.

Chapter 6

Conclusion

We studied how to devise algorithms amenable to deep reinforcement learning in two-player zero-sum extensive-form games. Such algorithms have to enjoy (last/best) iterate convergence, learn with rolling trajectories, and avoid importance sampling. In Chapter 3 and Chapter 4, we show how to extend OMD and CFR variants individually to iterate convergence under full-information feedback. The iterate convergence is achieved by adding additional regularization to the original bilinear objective function. In Chapter 5, `QFR` is proposed and achieves all three key properties simultaneously. A key contribution of `QFR` is utilizing different learning rates in different infosets to learn with trajectory Q-values effectively instead of counterfactual values. Furthermore, to avoid importance sampling, we propose the *bidilated regularizer* for EFGs, which weights the local regularizer in each infosets by the reach probability of both players.

An interesting future direction is to scale up the algorithm presented in this thesis in large games, such as dark chess. Moreover, deriving algorithms with the same learning rates in all infosets is crucial,[1] as updates to neural network parameters do not support different learning rates across various infosets in large-scale games, where policies are approximated by neural networks.

[1]We found that `QFR` can learn with the same learning rates empirically, but it remains open to find an algorithm that soundly avoids different learning rates.

Appendix A

Omitted Proofs of Chapter 3

Lemma 3.0.4. Suppose that $\psi^{\mathcal{C}}$ is a 1-strongly convex function with respect to p-norm in convex set $\mathcal{C}$, such that $D_{\psi^{\mathcal{C}}}(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(0)}) \geq \frac{1}{2}\|\boldsymbol{x}^{(1)} - \boldsymbol{x}^{(0)}\|_p^2$ for any $\boldsymbol{x}^{(0)}, \boldsymbol{x}^{(1)} \in \mathcal{C}$. Then, for any $\boldsymbol{x}^{(0)}, \boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)} \in \mathcal{C}$ and $\tau_0 \geq 0$ satisfying,

$$\begin{aligned}\boldsymbol{x}^{(1)} &= \operatorname*{argmin}_{\boldsymbol{x}\in\mathcal{C}} \left\langle \boldsymbol{x}, \boldsymbol{g}^{(1)} \right\rangle + \tau_0 \nabla\psi^{\mathcal{C}}(\boldsymbol{x}) + D_{\psi^{\mathcal{C}}}(\boldsymbol{x}, \boldsymbol{x}^{(0)}) \\ \boldsymbol{x}^{(2)} &= \operatorname*{argmin}_{\boldsymbol{x}\in\mathcal{C}} \left\langle \boldsymbol{x}, \boldsymbol{g}^{(2)} \right\rangle + \tau_0 \nabla\psi^{\mathcal{C}}(\boldsymbol{x}) + D_{\psi^{\mathcal{C}}}(\boldsymbol{x}, \boldsymbol{x}^{(0)}),\end{aligned}$$

we have

$$\left\|\boldsymbol{x}^{(1)} - \boldsymbol{x}^{(2)}\right\|_p \leq \frac{1}{1+\tau_0} \left\|\boldsymbol{g}^{(1)} - \boldsymbol{g}^{(2)}\right\|_q, \tag{3.0.4}$$

where $q \geq 1$ and $\frac{1}{p} + \frac{1}{q} = 1$.

Proof. By the first-order optimality of $\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}$, we have

$$\begin{aligned}\left\langle \boldsymbol{g}^{(1)} + (1+\tau_0)\nabla\psi^{\mathcal{C}}(\boldsymbol{x}^{(1)}) - \nabla\psi^{\mathcal{C}}(\boldsymbol{x}^{(0)}), \boldsymbol{x}^{(2)} - \boldsymbol{x}^{(1)} \right\rangle \geq 0 \\ \left\langle \boldsymbol{g}^{(2)} + (1+\tau_0)\nabla\psi^{\mathcal{C}}(\boldsymbol{x}^{(2)}) - \nabla\psi^{\mathcal{C}}(\boldsymbol{x}^{(0)}), \boldsymbol{x}^{(1)} - \boldsymbol{x}^{(2)} \right\rangle \geq 0.\end{aligned}$$

Summing up and rearranging the terms,

$$\left\langle \boldsymbol{x}^{(2)} - \boldsymbol{x}^{(1)}, \boldsymbol{g}^{(1)} - \boldsymbol{g}^{(2)} \right\rangle \geq (1+\tau_0) \left\langle \nabla\psi^{\mathcal{C}}(\boldsymbol{x}^{(1)}) - \nabla\psi^{\mathcal{C}}(\boldsymbol{x}^{(2)}), \boldsymbol{x}^{(1)} - \boldsymbol{x}^{(2)} \right\rangle. \tag{A.0.1}$$

To bound the right-hand-side of the inequality above, by strong convexity of $\psi^{\mathcal{C}}$, we have

$$\begin{aligned}\left\langle \nabla\psi^{\mathcal{C}}(\boldsymbol{x}^{(1)}), \boldsymbol{x}^{(1)} - \boldsymbol{x}^{(2)} \right\rangle \geq \psi^{\mathcal{C}}(\boldsymbol{x}^{(1)}) - \psi^{\mathcal{C}}(\boldsymbol{x}^{(2)}) + \frac{1}{2}\left\|\boldsymbol{x}^{(1)} - \boldsymbol{x}^{(2)}\right\|_p^2, \\ \left\langle \nabla\psi^{\mathcal{C}}(\boldsymbol{x}^{(2)}), \boldsymbol{x}^{(2)} - \boldsymbol{x}^{(1)} \right\rangle \geq \psi^{\mathcal{C}}(\boldsymbol{x}^{(2)}) - \psi^{\mathcal{C}}(\boldsymbol{x}^{(1)}) + \frac{1}{2}\left\|\boldsymbol{x}^{(1)} - \boldsymbol{x}^{(2)}\right\|_p^2.\end{aligned}$$

Summing them up we have

$$\left\langle \nabla\psi^{\mathcal{C}}(\boldsymbol{x}^{(1)}) - \nabla\psi^{\mathcal{C}}(\boldsymbol{x}^{(2)}), \boldsymbol{x}^{(1)} - \boldsymbol{x}^{(2)} \right\rangle \geq \|\boldsymbol{x}^{(1)} - \boldsymbol{x}^{(2)}\|_p^2.$$

Combining with (A.0.1),

$$\left\langle \boldsymbol{x}^{(2)} - \boldsymbol{x}^{(1)}, \boldsymbol{g}^{(1)} - \boldsymbol{g}^{(2)} \right\rangle \geq (1+\tau_0)\left\|\boldsymbol{x}^{(1)} - \boldsymbol{x}^{(2)}\right\|_p^2. \tag{A.0.2}$$

Finally, by Hölder's Inequality,

$$\left\langle \boldsymbol{x}^{(2)} - \boldsymbol{x}^{(1)}, \boldsymbol{g}^{(1)} - \boldsymbol{g}^{(2)} \right\rangle \le \left\| \boldsymbol{x}^{(1)} - \boldsymbol{x}^{(2)} \right\|_p \cdot \left\| \boldsymbol{g}^{(1)} - \boldsymbol{g}^{(2)} \right\|_q,$$

and as a result $\left\| \boldsymbol{x}^{(1)} - \boldsymbol{x}^{(2)} \right\|_p \le \frac{1}{1+\tau_0} \left\| \boldsymbol{g}^{(1)} - \boldsymbol{g}^{(2)} \right\|_q$ as claimed. □

Lemma 3.0.5. For any $\tau_0 \ge 0$ and sequence-form strategy profile μ, we have

$$\tau_0 \psi^{\Pi}(\mu) - \tau_0 \psi^{\Pi}(\mu^{\tau_0,*}) + \langle F(\mu), \mu - \mu^{\tau_0,*} \rangle \ge \tau_0 D_{\psi^{\Pi}}(\mu, \mu^{\tau_0,*}). \tag{3.0.5}$$

Proof. By definition of NE, we have (argmin is unique when $\tau_0 > 0$)

$$\mu_1^{\tau_0,*} \in \operatorname*{argmin}_{\mu_1 \in \Pi_1} \langle \boldsymbol{U} \mu_2^{\tau_0,*}, \mu_1 \rangle + \tau_0 \psi^{\Pi_1}(\mu_1).$$

By first-order optimality, for any $\mu_1 \in \Pi_1$, we have

$$\langle \boldsymbol{U} \mu_2^{\tau_0,*}, \mu_1 - \mu_1^{\tau_0,*} \rangle + \tau_0 \left\langle \nabla \psi^{\Pi_1}(\mu_1^{\tau_0,*}), \mu_1 - \mu_1^{\tau_0,*} \right\rangle \ge 0.$$

Therefore,

$$\begin{aligned}
&\langle \boldsymbol{U} \mu_2^{\tau_0,*}, \mu_1 - \mu_1^{\tau_0,*} \rangle + \tau_0 \psi^{\Pi_1}(\mu_1) - \tau_0 \psi^{\Pi_1}(\mu_1^{\tau_0,*}) \\
=&\langle \boldsymbol{U} \mu_2^{\tau_0,*}, \mu_1 - \mu_1^{\tau_0,*} \rangle + \tau_0 D_{\psi^{\Pi_1}}(\mu_1, \mu_1^{\tau_0,*}) + \tau_0 \left\langle \nabla \psi^{\Pi_1}(\mu_1^{\tau_0,*}), \mu_1 - \mu_1^{\tau_0,*} \right\rangle \\
\ge&\tau_0 D_{\psi^{\Pi_1}}(\mu_1, \mu_1^{\tau_0,*}).
\end{aligned}$$

By summing the inequality above and its counterpart for $\mu_2, \mu_2^{\tau_0,*}$,

$$\langle F(\mu^{\tau_0,*}), \mu - \mu^{\tau_0,*} \rangle + \tau_0 \psi^{\Pi}(\mu) - \tau_0 \psi^{\Pi}(\mu^{\tau_0,*}) \ge \tau_0 D_{\psi^{\Pi}}(\mu, \mu^{\tau_0,*}).$$

By definition of F, we have $\langle F(\mu), \mu \rangle = 0$ for any $\mu \in \Pi$. Therefore,

$$\begin{aligned}
&\tau_0 \psi^{\Pi}(\mu) - \tau_0 \psi^{\Pi}(\mu^{\tau_0,*}) + \langle F(\mu), \mu - \mu^{\tau_0,*} \rangle \\
=&\tau_0 \psi^{\Pi}(\mu) - \tau_0 \psi^{\Pi}(\mu^{\tau_0,*}) + \langle F(\mu^{\tau_0,*}), \mu - \mu^{\tau_0,*} \rangle \ge \tau_0 D_{\psi^{\Pi}}(\mu, \mu^{\tau_0,*}).
\end{aligned}$$

□

A.1 Omitted Proof of Section 3.1

Lemma 3.1.1. For any $\tau > 0$ and sequence-form strategy profile $\mu^{\boldsymbol{\pi}} \in \Pi$, we have

$$\max_{\mu^{\widehat{\boldsymbol{\pi}}} \in \Pi} \left\langle F(\mu^{\boldsymbol{\pi}}), \mu^{\boldsymbol{\pi}} - \mu^{\widehat{\boldsymbol{\pi}}} \right\rangle \le \tau \max_{\mu^{\boldsymbol{\pi}'} \in \Pi} \psi^{\Pi}(\mu^{\boldsymbol{\pi}'}) + \sqrt{\left(\sum_{s \in \mathcal{S}} |\mathcal{A}_s| \right) D_{\psi^{\Pi}}(\mu^{\tau,*}, \mu^{\boldsymbol{\pi}})}. \tag{3.1.1}$$

Proof. For any $\mu^{\widehat{\boldsymbol{\pi}}} \in \Pi$, we have

$$\begin{aligned}
\left\langle F(\mu^{\boldsymbol{\pi}}), \mu^{\boldsymbol{\pi}} - \mu^{\widehat{\boldsymbol{\pi}}} \right\rangle =& \left\langle F(\mu^{\tau,*}), \mu^{\tau,*} - \mu^{\widehat{\boldsymbol{\pi}}} \right\rangle + \tau \psi^{\Pi}(\mu^{\tau,*}) - \tau \psi^{\Pi}(\mu^{\widehat{\boldsymbol{\pi}}}) \\
&- \tau \psi^{\Pi}(\mu^{\tau,*}) + \tau \psi^{\Pi}(\mu^{\widehat{\boldsymbol{\pi}}}) + \left\langle F(\mu^{\widehat{\boldsymbol{\pi}}}), \mu^{\boldsymbol{\pi}} - \mu^{\tau,*} \right\rangle \\
\overset{(i)}{\le}& 0 + \tau \max_{\mu^{\boldsymbol{\pi}'} \in \Pi} \psi^{\Pi}(\mu^{\boldsymbol{\pi}'}) + \left\| F(\mu^{\widehat{\boldsymbol{\pi}}}) \right\|_\infty \cdot \left\| \mu^{\boldsymbol{\pi}} - \mu^{\tau,*} \right\|_1.
\end{aligned}$$

where (i) is by the definition of $\mu^{\tau,*}$, the non-negativity of ψ^{Π}, and Hölder's Inequality. Then, since we assume the utility $\mathcal{U}_i(h)$ of player $i \in [2]$ at any node $h \in \mathcal{H}$ is bounded in $[-1,1]$, we have $\left|(\mu_1^{\pi_1})^\top \boldsymbol{U} \mu_2^{\pi_2}\right| \leq 1$ for any sequence-form strategy profile $\mu^{\boldsymbol{\pi}}$. Moreover, since player 1 can play deterministically to reach any infoset $s \in \mathcal{S}_1$ and then choose action $a \in \mathcal{A}_s$ deterministically, $\|\boldsymbol{U}\mu_2^{\pi_2}\|_\infty \leq \max_{\mu_1^{\widehat{\pi}_1} \in \Pi_1} \left|\left(\mu_1^{\widehat{\pi}_1}\right)^\top \boldsymbol{U}\mu_2^{\pi_2}\right| \leq 1$. Similarly, $\left\|\boldsymbol{U}^\top \mu_1^{\pi_1}\right\|_\infty \leq 1$. Therefore, $\left\|F(\mu^{\widehat{\boldsymbol{\pi}}})\right\|_\infty \leq 1$, and we have

$$\begin{aligned}
\left\langle F(\mu^{\boldsymbol{\pi}}), \mu^{\boldsymbol{\pi}} - \mu^{\widehat{\boldsymbol{\pi}}} \right\rangle \leq & \tau \max_{\mu^{\boldsymbol{\pi}'} \in \Pi} \psi^{\Pi}(\mu^{\boldsymbol{\pi}'}) + \|\mu^{\boldsymbol{\pi}} - \mu^{\tau,*}\|_1 \\
\leq & \tau \max_{\mu^{\boldsymbol{\pi}'} \in \Pi} \psi^{\Pi}(\mu^{\boldsymbol{\pi}'}) + \sqrt{\sum_{s \in \mathcal{S}} |\mathcal{A}_s|}\, \|\mu^{\boldsymbol{\pi}} - \mu^{\tau,*}\| \\
\overset{(i)}{\leq} & \tau \max_{\mu^{\boldsymbol{\pi}'} \in \Pi} \psi^{\Pi}(\mu^{\boldsymbol{\pi}'}) + \sqrt{\left(\sum_{s \in \mathcal{S}} |\mathcal{A}_s|\right) D_{\psi^{\Pi}}(\mu^{\tau,*}, \mu^{\boldsymbol{\pi}})}.
\end{aligned}$$

(i) is by the 1-strong convexity of ψ^{Π}. □

Theorem 3.1.2. With the shrinking algorithm described above, the duality gap satisfies $\max_{\mu^{\widehat{\boldsymbol{\pi}}} \in \Pi} \left\langle F(\bar{\mu}^{(t+1)}, \bar{\mu}^{(t+1)} - \mu^{\widehat{\boldsymbol{\pi}}} \right\rangle \leq \widetilde{\mathcal{O}}\left(\frac{1}{t}\right)$ for $t = 1, 2, ..., T$. Moreover, let Π^* be the set of NE of the original game, the distance to the set of NE also satisfies $\left\|\bar{\mu}^{(t+1)} - \mathrm{Proj}_{\Pi^*}\left(\bar{\mu}^{(t+1)}\right)\right\| \leq \widetilde{\mathcal{O}}\left(\frac{1}{t}\right)$.

Proof. To ensure the following holds for some $\epsilon > 0$,

$$\sqrt{\left(\sum_{s \in \mathcal{S}} |\mathcal{A}_s|\right) D_{\psi^{\Pi}}\left(\mu^{\tau,*}, \bar{\mu}^{(T+1)}\right)} \leq \frac{\epsilon}{2},$$

the following need to hold,

$$\left(\frac{1}{1+\eta\tau}\right)^T D_{\psi^{\Pi}}\left(\mu^{\tau,*}, \bar{\mu}^{(1)}\right) \leq \frac{\epsilon^2}{4\sum_{s \in \mathcal{S}} |\mathcal{A}_s|}.$$

according to Theorem 3.0.2. Moreover, since $1 + x \leq e^x$ for any $x \in \mathbb{R}$, we have

$$\begin{aligned}
\left(\frac{1}{1+\eta\tau}\right)^T D_{\psi^{\Pi}}\left(\mu^{\tau,*}, \bar{\mu}^{(1)}\right) &= \left(1 - \frac{\eta\tau}{1+\eta\tau}\right)^T D_{\psi^{\Pi}}\left(\mu^{\tau,*}, \bar{\mu}^{(1)}\right) \\
&\leq \exp\left(-\frac{\eta\tau}{1+\eta\tau} T\right) D_{\psi^{\Pi}}\left(\mu^{\tau,*}, \bar{\mu}^{(1)}\right).
\end{aligned}$$

Therefore, it holds by letting $\eta = \frac{1}{2L}$, $\tau = \frac{\epsilon}{2\max_{\mu^{\boldsymbol{\pi}'} \in \Pi} \psi^{\Pi}(\mu^{\boldsymbol{\pi}'})}$, and

$$T = \left\lceil (1+\eta\tau) \frac{2\log 2 + \log \sum_{s \in \mathcal{S}} |\mathcal{A}_s| + \log D_{\psi^{\Pi}}\left(\mu^{\tau,*}, \bar{\mu}^{(1)}\right) + 2\log\frac{1}{\epsilon}}{\eta\tau} \right\rceil = \widetilde{\mathcal{O}}\left(\frac{1}{\epsilon}\right).$$

Furthermore, by Lemma 3.1.1,

$$\max_{\mu^{\widehat{\boldsymbol{\pi}}} \in \Pi} \left\langle F(\mu^{\boldsymbol{\pi}}), \bar{\mu}^{(T+1)} - \mu^{\widehat{\boldsymbol{\pi}}} \right\rangle \leq \epsilon.$$

Therefore, in the $K+1$ episode, *i.e.*, $\tau = \frac{\tau_0}{2^K}$, it takes $\widetilde{\mathcal{O}}\left(\frac{1}{\tau}\right)$ timesteps to ensure the duality is lower than τ. Hence, for any $\epsilon \in [\frac{\tau_0}{2^K}, \frac{\tau_0}{2^{K-1}})$ $(K > 0)$, it takes $K+1$ episode to ensure the duality gap is lower than ϵ. The total number of timesteps performed so far is no larger than

$$\sum_{k=0}^{K} \widetilde{\mathcal{O}}\left(\frac{2^k}{\tau_0}\right) \leq \widetilde{\mathcal{O}}(\frac{1}{\epsilon}). \qquad \square$$

Appendix B

Omitted Proofs of Chapter 4

Lemma 4.0.1. For any sequence of sequence-form strategy profiles $\mu^{(1)}, \mu^{(2)}, ..., \mu^{(T)}$ with length $T > 0$ and sequence-form strategy $\mu_i^{\pi_i} \in \Pi_i$ of player $i \in [2]$, we have

$$G_i^{\psi^\Pi,(T)}(\mu_i^{\pi_i}) = \sum_{s \in \mathcal{S}_i} \mu_i^{\pi_i}(\sigma(s)) G_s^{\psi^\Pi,(T)}(\pi_i) \tag{4.0.2}$$

$$R_i^{\psi^\Pi,(T)} \leq \max_{\mu_i^{\widehat{\pi}_i} \in \Pi_i} \sum_{s \in \mathcal{S}_i} \mu_i^{\widehat{\pi}_i}(\sigma(s)) R_s^{\psi^\Pi,(T)}. \tag{4.0.3}$$

Proof. We define the scalar *subtree value* $\mathrm{Sub}_s^{(t)}(\mu)$ recursively. For any infoset $s \in \mathcal{S}_1$ (similarly for any $s \in \mathcal{S}_2$), let $(\boldsymbol{U}\mu_2^{(t)})_{(s,a)}$ be the value of vector $\boldsymbol{U}\mu_2^{(t)}$ indexed by (s,a), then we have

$$\mathrm{Sub}_s^{(t)}(\mu_1^{\pi_1}) \coloneqq \sum_{a \in \mathcal{A}_s} \pi_1(a \,|\, s) \left((\boldsymbol{U}\mu_2^{(t)})_{(s,a)} + \sum_{s' \in \mathcal{S}_1 \,:\, (s,a) \sqsubseteq s'} \mathrm{Sub}_{s'}^{(t)}(\mu_1^{\pi_1}) \right) + \tau \psi_s^\Delta(\pi_1(\cdot \,|\, s)).$$

By definition, for any $\mu_1^{\pi_1} \in \Pi_1$, we have

$$\begin{aligned} G_1^{\psi^\Pi,(T)}(\mu_1^{\pi_1}) &= \sum_{t=1}^{T} \left\langle \boldsymbol{U}\mu_2^{(t)}, \mu_1^{(t)} - \mu_1^{\pi_1} \right\rangle + \tau\psi^{\Pi_1}(\mu_1^{(t)}) - \tau\psi^{\Pi_1}(\mu_1^{\pi_1}) \\ &= \sum_{t=1}^{T} \sum_{s \in \mathcal{S}_1 \,:\, \sigma(s) = \emptyset} \left(\mathrm{Sub}_s^{(t)}(\mu_1^{(t)}) - \mathrm{Sub}_s^{(t)}(\mu_1^{\pi_1}) \right) \\ &= \sum_{s \in \mathcal{S}_1 \,:\, \sigma(s) = \emptyset} \left(\mathrm{Sub}_s^{(t)}(\mu_1^{(t)}) - \mathrm{Sub}_s^{(t)}(\mu_1^{\pi_1}) \right) \end{aligned}$$

Then,

$$
\begin{aligned}
&\sum_{t=1}^{T}\left(\mathrm{Sub}_s^{(t)}(\mu_1^{(t)})-\mathrm{Sub}_s^{(t)}(\mu_1^{\pi_1})\right)\\
=&\sum_{t=1}^{T}\mathrm{Sub}_s^{(t)}(\mu_1^{(t)})-\sum_{t=1}^{T}\left(\tau\psi_s^{\Delta}(\pi_1(\cdot\,|\,s))+\sum_{a\in\mathcal{A}_s}\pi_1(a\,|\,s)(\boldsymbol{U}\mu_2^{(t)})_{(s,a)}\right)\\
&-\sum_{a\in\mathcal{A}_s}\pi_1(a\,|\,s)\sum_{s'\in\mathcal{S}_1\colon(s,a)\sqsubseteq s'}\sum_{t=1}^{T}\mathrm{Sub}_{s'}^{(t)}(\mu_1^{\pi_1})\\
=&\sum_{t=1}^{T}\mathrm{Sub}_s^{(t)}(\mu_1^{(t)})-\sum_{t=1}^{T}\left(\tau\psi_s^{\Delta}(\pi_1(\cdot\,|\,s))+\sum_{a\in\mathcal{A}_s}\pi_1(a\,|\,s)(\boldsymbol{U}\mu_2^{(t)})_{(s,a)}\right)\\
&-\sum_{a\in\mathcal{A}_s}\pi_1(a\,|\,s)\sum_{s'\in\mathcal{S}_1\colon(s,a)\sqsubseteq s'}\sum_{t=1}^{T}\left(\mathrm{Sub}_{s'}^{(t)}(\mu_1^{(t)})-\left(\mathrm{Sub}_{s'}^{(t)}(\mu_1^{(t)})-\mathrm{Sub}_{s'}^{(t)}(\mu_1^{\pi_1})\right)\right)\\
=&\sum_{t=1}^{T}\mathrm{Sub}_s^{(t)}(\mu_1^{(t)})\\
&-\sum_{t=1}^{T}\left(\tau\psi_s^{\Delta}(\pi_1(\cdot\,|\,s))+\sum_{a\in\mathcal{A}_s}\pi_1(a\,|\,s)\left((\boldsymbol{U}\mu_2^{(t)})_{(s,a)}+\sum_{s'\in\mathcal{S}_1\colon(s,a)\sqsubseteq s'}\sum_{t=1}^{T}\mathrm{Sub}_{s'}^{(t)}(\mu_1^{(t)})\right)\right)\\
&+\sum_{a\in\mathcal{A}_s}\pi_1(a\,|\,s)\sum_{t=1}^{T}\left(\mathrm{Sub}_{s'}^{(t)}(\mu_1^{(t)})-\mathrm{Sub}_{s'}^{(t)}(\mu_1^{\pi_1})\right).
\end{aligned}
$$

By letting $\boldsymbol{\pi}^{(t)}$ be the behavior-form strategy profile corresponding to $\mu^{(t)}$, we have

$$
\begin{aligned}
&\sum_{t=1}^{T}\left(\mathrm{Sub}_s^{(t)}(\mu_1^{(t)})-\mathrm{Sub}_s^{(t)}(\mu_1^{\pi_1})\right)\\
=&\sum_{t=1}^{T}\left(\left\langle \mathrm{CF}_1^{\psi^{\Pi},\boldsymbol{\pi}^{(t)}}(s,\cdot),\pi_1^{(t)}(\cdot\,|\,s)-\pi_1(\cdot\,|\,s)\right\rangle+\tau\psi_s^{\Delta}(\pi_1^{(t)}(\cdot\,|\,s))-\tau\psi_s^{\Delta}(\pi_1(\cdot\,|\,s))\right)\\
&+\sum_{a\in\mathcal{A}_s}\pi_1(a\,|\,s)\sum_{t=1}^{T}\left(\mathrm{Sub}_{s'}^{(t)}(\mu_1^{(t)})-\mathrm{Sub}_{s'}^{(t)}(\mu_1^{\pi_1})\right)\\
=&G_s^{\psi^{\Pi},(T)}(\pi_1)+\sum_{a\in\mathcal{A}_s}\pi_1(a\,|\,s)\sum_{t=1}^{T}\left(\mathrm{Sub}_{s'}^{(t)}(\mu_1^{(t)})-\mathrm{Sub}_{s'}^{(t)}(\mu_1^{\pi_1})\right).
\end{aligned}
$$

By applying it recursively, we will get for any $\mu_1^{\pi_1}\in\Pi_1$,

$$
G_1^{\psi^{\Pi},(T)}(\mu_1^{\pi_1})=\sum_{s\in\mathcal{S}_1}\mu_1^{\pi_1}(\sigma(s))G_s^{\psi^{\Pi},(T)}(\pi_1),\tag{B.0.1}
$$

which completes the first half of the proof.

By (B.0.1), we have

$$\begin{aligned} R_1^{\psi^\Pi,(T)} = \max_{\mu_1^{\widehat{\pi}_1} \in \Pi_1} G_1^{\psi^\Pi,(T)}(\mu_1^{\widehat{\pi}_1}) &= \max_{\mu_1^{\widehat{\pi}_1} \in \Pi_1} \sum_{s \in \mathcal{S}_1} \mu_1^{\widehat{\pi}_1}(\sigma(s)) G_s^{\psi^\Pi,(T)}(\widehat{\pi}_1) \\ &\leq \max_{\mu_1^{\widehat{\pi}_1} \in \Pi_1} \sum_{s \in \mathcal{S}_1} \mu_1^{\widehat{\pi}_1}(\sigma(s)) \max_{\widehat{\pi}'_1} G_s^{\psi^\Pi,(T)}(\widehat{\pi}'_1) \\ &= \max_{\mu_1^{\widehat{\pi}_1} \in \Pi_1} \sum_{s \in \mathcal{S}_1} \mu_1^{\widehat{\pi}_1}(\sigma(s)) R_s^{\psi^\Pi,(T)}, \end{aligned}$$

which completes the proof. The proof for player 2 is similar. □

Appendix C

Omitted Proofs of Chapter 5

In this chapter, we introduce the missing proofs of Chapter 5.

C.1 Bidilated Regularizer

Dilated regularizer [Hoda et al., 2010] is the foundation of previous work [Lee et al., 2021, Liu et al., 2023, Sokota et al., 2023] to apply mirror-descent and its variants on sequence-form strategies. Recently, additional regularization has become a powerful tool for learning in EFGs [Liu et al., 2023, Sokota et al., 2023]. Specifically, we can change the objective of the game to $\max_{\mu_1^{\pi_1} \in \Pi_1} \min_{\mu_2^{\pi_2} \in \Pi_2} (\mu_1^{\pi_1})^\top \boldsymbol{A} \mu_2^{\pi_2} - \tau \psi^{\Pi_1}(\mu_1^{\pi_1}) + \tau \psi^{\Pi_2}(\mu_2^{\pi_2})$, where $\tau \psi^{\Pi_1}(\mu_1^{\pi_1}), \tau \psi^{\Pi_2}(\mu_2^{\pi_2})$ is the additional regularizer and τ controls its magnitude. By adding the additional regularizer, the objective becomes strongly convex-concave instead of convex-concave, and thus linear convergence rate can be achieved.

However, the dilated regularizer of player $i \in [2]$ is $\psi^{\Pi_i}(\mu_i^{\pi_i}) = \sum_{s \in \mathcal{S}_i} \mu_i^{\pi_i}(\sigma(s)) \psi_s^{\Delta}(\pi_i(\cdot \mid s))$, which only counts the reach probability $\mu_i^{\pi_i}(\sigma(s))$ of player i. Therefore, when sampling a trajectory, to estimate the additional regularization, importance sampling is needed to offset the reach probability of player $3 - i$ and the chance player, which causes a large dispersion of feedback. Therefore, to avoid importance sampling on the regularizer, we propose the bidilated regularizer in this section, to which all players contribute symmetrically. The bidilated regularizer of player 1 is defined as,

$$\psi_{\mathrm{bi}}^{\Pi_1}(\mu_1^{\pi_1}, \mu_2^{\pi_2}) \coloneqq \sum_{s \in \mathcal{S}_1} \mu_1^{\pi_1}(\sigma(s)) \left(\sum_{h \in s} \mu_c(h) \mu_2^{\pi_2}(\sigma_2(h)) \right) \psi_s^{\Delta}(\pi_1(\cdot \mid s)). \tag{C.1.1}$$

The additional term is the probability of reaching infoset s contributed by player 2 and the chance player. The bidilated regularizer for player 2 can also be defined similarly. In the following, we will show that several preferable properties of dilated regularizer still hold for its bidilated version.

Firstly, the bidilated regularizer $\psi_{\mathrm{bi}}^{\Pi_1}(\mu_1^{\pi_1}, \mu_2^{\pi_2})$ is still convex with respect to $\mu_1^{\pi_1}$ and $\mu_2^{\pi_2}$ individually. This can be inferred from the fact that the dilated regularizer is convex with respect to $\mu_1^{\pi_1}$ [Hoda et al., 2010]. By enforcing $\pi_2(\cdot \mid s) \geq \gamma_s \boldsymbol{\nu}_s$ for every $s \in \mathcal{S}_2$ with $\gamma_s > 0, \boldsymbol{\nu}_s \in \Delta^{|\mathcal{A}_s|}$, where $\boldsymbol{\nu}_s$ has full support, we have $\mu_2^{\pi_2}(s, a) \geq \gamma > 0$ for any $s \in \mathcal{S}_2, a \in \mathcal{A}_s$, where γ is a constant. Then, we have the following lemma.

Lemma C.1.1. For any $\tau, \gamma > 0$, the Nash equilibrium $\boldsymbol{\mu}^{(\tau,\gamma),*} = (\mu_1^{(\tau,\gamma),*}, \mu_2^{(\tau,\gamma),*})$ of (5.0.1) is unique.

Proof. Let define $F_1^\tau(\mu_1^{\pi_1}, \mu_2^{\pi_2}) \coloneqq -\boldsymbol{A}\mu_2^{\pi_2} + \tau\nabla_{\mu_1^{\pi_1}}\psi_{\text{bi}}^{\Pi_1}(\mu_1^{\pi_1}, \mu_2^{\pi_2}) - \tau\nabla_{\mu_1^{\pi_1}}\psi_{\text{bi}}^{\Pi_2}(\mu_1^{\pi_1}, \mu_2^{\pi_2})$ and $F_2^\tau(\mu_1^{\pi_1}, \mu_2^{\pi_2}) \coloneqq \boldsymbol{A}^\top\mu_1^{\pi_1} + \tau\nabla_{\mu_2^{\pi_2}}\psi_{\text{bi}}^{\Pi_2}(\mu_1^{\pi_1}, \mu_2^{\pi_2}) - \tau\nabla_{\mu_2^{\pi_2}}\psi_{\text{bi}}^{\Pi_2}(\mu_1^{\pi_1}, \mu_2^{\pi_2})$.

For any $\mu_1^{\pi_1'} \in \Pi_1$, we have

$$
\begin{aligned}
&\left\langle F_1^\tau(\mu_1^{\pi_1}, \mu_2^{\pi_2}), \mu_1^{\pi_1} - \mu_1^{\pi_1'} \right\rangle \\
=&\left\langle -\boldsymbol{A}\mu_2^{\pi_2} + \tau\nabla_{\mu_1^{\pi_1}}\psi_{\text{bi}}^{\Pi_1}(\mu_1^{\pi_1}, \mu_2^{\pi_2}) - \tau\nabla_{\mu_1^{\pi_1}}\psi_{\text{bi}}^{\Pi_2}(\mu_1^{\pi_1}, \mu_2^{\pi_2}), \mu_1^{\pi_1} - \mu_1^{\pi_1'} \right\rangle \\
=&-\left\langle \boldsymbol{A}\mu_2^{\pi_2}, \mu_1^{\pi_1} - \mu_1^{\pi_1'} \right\rangle + \tau D_{\psi_{\text{bi}}^{\Pi_1}(\cdot,\mu_2^{\pi_2})}\left(\mu_1^{\pi_1'}, \mu_1^{\pi_1}\right) \\
&- \tau\psi_{\text{bi}}^{\Pi_1}(\mu_1^{\pi_1'}, \mu_2^{\pi_2}) + \tau\psi_{\text{bi}}^{\Pi_1}(\mu_1^{\pi_1}, \mu_2^{\pi_2}) - \tau\psi_{\text{bi}}^{\Pi_2}(\mu_1^{\pi_1}, \mu_2^{\pi_2}) + \tau\psi_{\text{bi}}^{\Pi_2}(\mu_1^{\pi_1'}, \mu_2^{\pi_2}).
\end{aligned}
$$

The last line uses the fact that $\psi_{\text{bi}}^{\Pi_2}(\mu_1^{\pi_1}, \mu_2^{\pi_2})$ is linear with respect to $\mu_1^{\pi_1}$.

The counterpart of $\mu_2^{\pi_2}$ is

$$
\begin{aligned}
&\left\langle F_2^\tau(\mu_1^{\pi_1}, \mu_2^{\pi_2}), \mu_2^{\pi_2} - \mu_2^{\pi_2'} \right\rangle \\
=&\left\langle \boldsymbol{A}^\top\mu_1^{\pi_1} + \tau\nabla_{\mu_2^{\pi_2}}\psi_{\text{bi}}^{\Pi_2}(\mu_1^{\pi_1}, \mu_2^{\pi_2}) - \tau\nabla_{\mu_2^{\pi_2}}\psi_{\text{bi}}^{\Pi_2}(\mu_1^{\pi_1}, \mu_2^{\pi_2}), \mu_2^{\pi_2} - \mu_2^{\pi_2'} \right\rangle \\
=&\left\langle \boldsymbol{A}^\top\mu_1^{\pi_1}, \mu_2^{\pi_2} - \mu_2^{\pi_2'} \right\rangle + \tau D_{\psi_{\text{bi}}^{\Pi_2}(\mu_1^{\pi_1},\cdot)}\left(\mu_2^{\pi_2'}, \mu_2^{\pi_2}\right) \\
&- \tau\psi_{\text{bi}}^{\Pi_2}(\mu_1^{\pi_1}, \mu_2^{\pi_2'}) + \tau\psi_{\text{bi}}^{\Pi_2}(\mu_1^{\pi_1}, \mu_2^{\pi_2}) - \tau\psi_{\text{bi}}^{\Pi_1}(\mu_1^{\pi_1}, \mu_2^{\pi_2}) + \tau\psi_{\text{bi}}^{\Pi_1}(\mu_1^{\pi_1}, \mu_2^{\pi_2'}).
\end{aligned}
$$

Let $\boldsymbol{\mu}^{\boldsymbol{\pi}} = (\mu_1^{\pi_1}, \mu_2^{\pi_2})$ and $F^\tau(\boldsymbol{\mu}^{\boldsymbol{\pi}}) = (F_1^\tau(\mu_1^{\pi_1}, \mu_2^{\pi_2}), F_2^\tau(\mu_1^{\pi_1}, \mu_2^{\pi_2}))$. Then by taking the summation of equations above, we have

$$
\begin{aligned}
&\left\langle F^\tau(\boldsymbol{\mu}^{\boldsymbol{\pi}}), \boldsymbol{\mu}^{\boldsymbol{\pi}} - \boldsymbol{\mu}^{\boldsymbol{\pi}'} \right\rangle \\
=&-(\mu_1^{\pi_1})^\top\boldsymbol{A}\mu_2^{\pi_2'} + (\mu_1^{\pi_1'})^\top\boldsymbol{A}\mu_2^{\pi_2} + \tau D_{\psi_{\text{bi}}^{\Pi_1}(\cdot,\mu_2^{\pi_2})}\left(\mu_1^{\pi_1'}, \mu_1^{\pi_1}\right) + \tau D_{\psi_{\text{bi}}^{\Pi_2}(\mu_1^{\pi_1},\cdot)}\left(\mu_2^{\pi_2'}, \mu_2^{\pi_2}\right) \\
&+ \tau\psi_{\text{bi}}^{\Pi_1}(\mu_1^{\pi_1}, \mu_2^{\pi_2'}) - \tau\psi_{\text{bi}}^{\Pi_1}(\mu_1^{\pi_1'}, \mu_2^{\pi_2}) + \tau\psi_{\text{bi}}^{\Pi_2}(\mu_1^{\pi_1'}, \mu_2^{\pi_2}) - \tau\psi_{\text{bi}}^{\Pi_2}(\mu_1^{\pi_1}, \mu_2^{\pi_2'}).
\end{aligned}
$$

Then,

$$
\begin{aligned}
&\left\langle F^\tau(\boldsymbol{\mu}^{\boldsymbol{\pi}}) - F^\tau(\boldsymbol{\mu}^{\boldsymbol{\pi}'}), \boldsymbol{\mu}^{\boldsymbol{\pi}} - \boldsymbol{\mu}^{\boldsymbol{\pi}'} \right\rangle \\
=&\tau\Big(D_{\psi_{\text{bi}}^{\Pi_1}(\cdot,\mu_2^{\pi_2})}\left(\mu_1^{\pi_1'}, \mu_1^{\pi_1}\right) + D_{\psi_{\text{bi}}^{\Pi_2}(\mu_1^{\pi_1},\cdot)}\left(\mu_2^{\pi_2'}, \mu_2^{\pi_2}\right) \\
&+ D_{\psi_{\text{bi}}^{\Pi_1}(\cdot,\mu_2^{\pi_2'})}\left(\mu_1^{\pi_1}, \mu_1^{\pi_1'}\right) + D_{\psi_{\text{bi}}^{\Pi_2}(\mu_1^{\pi_1'},\cdot)}\left(\mu_2^{\pi_2}, \mu_2^{\pi_2'}\right)\Big).
\end{aligned}
$$

Since $\boldsymbol{\mu}^{\boldsymbol{\pi}}, \boldsymbol{\mu}^{\boldsymbol{\pi}'} \succeq \gamma$, $D_{\psi_{\text{bi}}^{\Pi_1}(\cdot,\mu_2^{\pi_2'})}\left(\mu_1^{\pi_1}, \mu_1^{\pi_1'}\right) \geq \gamma \min_{h\in\mathcal{H}} \mu_c(h) D_{\psi^{\Pi_1}}\left(\mu_1^{\pi_1}, \mu_1^{\pi_1'}\right)$ by Lemma 5.1.1. Moreover, there exists $M > 0$ so that $D_{\psi^{\Pi_1}}\left(\mu_1^{\pi_1}, \mu_1^{\pi_1'}\right) \geq M\left\|\mu_1^{\pi_1} - \mu_1^{\pi_1'}\right\|^2$ according to Hoda et al. [2010], Lee et al. [2021]. Therefore, the NE is unique when $\tau, \gamma > 0$ by Rosen [1965]. □

C.2 Stability of Trajectory Q-value and Q-Value

In this section, we will show the stability of trajectory Q-value and Q-value, *i.e.,* proving Property 2 when the regularizer in each infoset $s \in \mathcal{S}$ can be written as,

$$\psi_s^{\Delta}(\boldsymbol{u}) = \begin{cases} \frac{\alpha_s}{2} \sum_{a \in \mathcal{A}_s} u_a^2 & \text{(Euclidean Norm)} \\ \alpha_s \left(\log |\mathcal{A}_s| + \sum_{a \in \mathcal{A}_s} u_a \log u_a \right) & \text{(Negative Entropy)}, \end{cases} \tag{C.2.1}$$

where $\alpha_s > 0$ is a state-dependent constant. We add $\log |\mathcal{A}_s|$ to the negative entropy to ensure the regularizer is always positive. Previous work [Kroer et al., 2020] chose specific α_s to ensure the dilated regularizer associated with ψ_s^{Δ} is 1-strongly convex. For generality of the result, we keep the α_s in the regularizer.

Due to the symmetry between two players, we will only prove that for $s \in \mathcal{S}_1$. Moreover, to stabilize trajectory Q-value and Q-value, the learning rate need to satisfy the following conditions.

(A) $\max_{h \in s} \sum_{(s',a') \sqsubseteq \sigma_i(h)} \eta_{s'} \le \eta_s$ for any $s \in \mathcal{S}, i \in [2]$

(B) $6\eta_s^{\text{anc}} \max_{s' \in \mathcal{S}} \left(\frac{2\|\boldsymbol{q}\|_\infty}{\alpha_{s'}} + \frac{\tau}{M_1} \log \frac{1}{\gamma} \right) \le 1$ for any $s \in \mathcal{S}$, where $\|\boldsymbol{q}\|_\infty \coloneqq \max_{t \in [T], s \in \mathcal{S}} \left\| q^{(t)}(s, \cdot) \right\|_\infty$ and its upperbound is given in Lemma C.5.1

(C) $\eta_s \left(2\|\boldsymbol{q}\|_\infty + \frac{\tau \alpha_s}{M_1} \log \frac{1}{\gamma} \right) \le 1$ for any $s \in \mathcal{S}$

(A) ensures that $\sum_{(s',a') \sqsubseteq \sigma(s)} \eta_{s'} \le 2\eta_s^{\text{anc}}$ for any $s \in \mathcal{S}$. **(B)**, **(C)** ensure that the update at each iteration will not change the strategy too much.

C.2.1 Stability of Trajectory Q-value

Lemma C.2.1 (Stability of $m_s^{(t)}$ under Euclidean regularizer)**.** Consider when $\psi_s^{\Delta}(\boldsymbol{u}) = \frac{\alpha_s}{2} \sum_{a \in \mathcal{A}_s} u_a^2$ for any $s \in \mathcal{S}$ and **(A)** is satisfied. For any $s \in \mathcal{S}$ and $t = 1, 2, \cdots, T$, when $m_s^{(t)}$ is the trajectory Q-value feedback, we have

$$C_s^- = \frac{6}{\gamma^2} \max_{s' \in \mathcal{S}} C_{s'}^{\text{diff}} \qquad C_s^/ = \frac{6}{\gamma^2 M_1} \max_{s' \in \mathcal{S}} C_{s'}^{\text{diff}}. \tag{C.2.2}$$

Proof. For trajectory Q-value feedback, for any $s \in \mathcal{S}_1$,

$$\begin{aligned} \left| m_s^{(t+1)} - m_s^{(t)} \right| &= \left| \frac{1}{\mu_1^{(t+1)}(\sigma(s))} - \frac{1}{\mu_1^{(t)}(\sigma(s))} \right| \\ &= \frac{\left| \mu_1^{(t+1)}(\sigma(s)) - \mu_1^{(t)}(\sigma(s)) \right|}{\mu_1^{(t+1)}(\sigma(s)) \mu_1^{(t)}(\sigma(s))} \\ &= \frac{\left| \prod_{(s',a') \sqsubseteq \sigma(s)} \pi_1^{(t+1)}(a' \mid s') - \prod_{(s',a') \sqsubseteq \sigma(s)} \pi_1^{(t)}(a' \mid s') \right|}{\mu_1^{(t+1)}(\sigma(s)) \mu_1^{(t)}(\sigma(s))} \\ &\le \frac{1}{\gamma^2} \sum_{(s',a') \sqsubseteq \sigma(s)} \left| \pi_1^{(t+1)}(a' \mid s') - \pi_1^{(t)}(a' \mid s') \right|. \end{aligned}$$

In the last line, we use the fact $\mu_1^{(t+1)}(\sigma(s)), \mu_1^{(t)}(\sigma(s)) \geq \gamma$, and

$$\begin{aligned}
&\left| \prod_{(s',a') \sqsubseteq \sigma(s)} \pi_1^{(t+1)}(a' \,|\, s') - \prod_{(s',a') \sqsubseteq \sigma(s)} \pi_1^{(t)}(a' \,|\, s') \right| \\
\leq & \prod_{(s',a') \sqsubseteq \sigma_1(\sigma(s))} \pi_1^{(t+1)}(a' \,|\, s') \left| \pi_1^{(t+1)}(\sigma(s)) - \pi_1^{(t)}(\sigma(s)) \right| \\
& + \pi_1^{(t)}(\sigma(s)) \left| \prod_{(s',a') \sqsubseteq \sigma(\sigma(s))} \pi_1^{(t+1)}(a' \,|\, s') - \prod_{(s',a') \sqsubseteq \sigma(\sigma(s))} \pi_1^{(t)}(a' \,|\, s') \right| \\
\leq & \left| \pi_1^{(t+1)}(\sigma(s)) - \pi_1^{(t)}(\sigma(s)) \right| + \left| \prod_{(s',a') \sqsubseteq \sigma(\sigma(s))} \pi_1^{(t+1)}(a' \,|\, s') - \prod_{(s',a') \sqsubseteq \sigma(\sigma(s))} \pi_1^{(t)}(a' \,|\, s') \right|.
\end{aligned}$$

We abuse the notion of $\pi_1(\sigma(s))$ as $\pi_1(a' \,|\, s')$ and $\sigma(\sigma(s)) = \sigma(s')$, given $\sigma(s) = (s', a')$.

By recursively applying the process above, we will get

$$\left| \prod_{(s',a') \sqsubseteq \sigma(s)} \pi_1^{(t+1)}(a' \,|\, s') - \prod_{(s',a') \sqsubseteq \sigma(s)} \pi_1^{(t)}(a' \,|\, s') \right| \leq \sum_{(s',a') \sqsubseteq \sigma(s)} \left| \pi_1^{(t+1)}(a' \,|\, s') - \pi_1^{(t)}(a' \,|\, s') \right|.$$

Lemma C.2.2. Consider the update-rule (5.1.2). When we choose ψ_s^Δ to be negative entropy or Euclidean distance, we have

$$\left\| \pi_{p(s)}^{(t)}(\cdot \,|\, s) - \bar{\pi}_{p(s)}^{(t)}(\cdot \,|\, s) \right\|_1, \left\| \bar{\pi}_{p(s)}^{(t+1)}(\cdot \,|\, s) - \bar{\pi}_{p(s)}^{(t)}(\cdot \,|\, s) \right\|_1 \leq C_s^{\mathrm{diff}} \eta_s, \tag{C.2.3}$$

where

$$C_s^{\mathrm{diff}} := \begin{cases} \frac{2}{\alpha_s} \left(2 \left\| \boldsymbol{q} \right\|_\infty + \frac{\tau \alpha_s}{M_1} \log \frac{1}{\gamma} \right) & \text{Negative Entropy} \\ \frac{|\mathcal{A}_s|}{\alpha_s} \left\| \boldsymbol{q} \right\|_\infty + \frac{2\sqrt{|\mathcal{A}_s|}\tau}{M_1} & \text{Euclidean Distance.} \end{cases} \tag{C.2.4}$$

The proof is postponed to Appendix C.5.3. By using Lemma C.2.2, we have

$$\begin{aligned}
& \left| m_s^{(t+1)} - m_s^{(t)} \right| \\
\leq & \frac{1}{\gamma^2} \sum_{(s',a') \sqsubseteq \sigma(s)} \left| \pi_1^{(t+1)}(a' \,|\, s') - \pi_1^{(t)}(a' \,|\, s') \right| \\
\leq & \frac{1}{\gamma^2} \sum_{(s',a') \sqsubseteq \sigma(s)} \Big(\left| \bar{\pi}_1^{(t+1)}(a' \,|\, s') - \bar{\pi}_1^{(t)}(a' \,|\, s') \right| + \left| \pi_1^{(t+1)}(a' \,|\, s') - \bar{\pi}_1^{(t+1)}(a' \,|\, s') \right| \\
& + \left| \bar{\pi}_1^{(t)}(a' \,|\, s') - \pi_1^{(t)}(a' \,|\, s') \right| \Big) \\
\leq & \frac{3}{\gamma^2} \sum_{(s',a') \sqsubseteq \sigma(s)} C_{s'}^{\mathrm{diff}} \eta_{s'}.
\end{aligned}$$

At the same time,

$$\left|\frac{m_s^{(t+1)}}{m_s^{(t)}}-1\right|=\frac{1}{m_s^{(t)}}\left|m_s^{(t+1)}-m_s^{(t)}\right|\le\frac{3\sum_{(s',a')\sqsubseteq\sigma(s)}C_{s'}^{\mathrm{diff}}\eta_{s'}}{\gamma^2 M_1}.$$

Therefore, $C_s^-=\frac{6}{\gamma^2}\max_{s'\in\mathcal{S}}C_{s'}^{\mathrm{diff}}$ and $C_s^{/}=\frac{6\max_{s'\in\mathcal{S}}C_{s'}^{\mathrm{diff}}}{\gamma^2 M_1}$ by (A). □

Lemma C.2.3 (Stability of $m_s^{(t)}$ under entropy regularizer)**.** Consider when $\psi_s^{\Delta}(\boldsymbol{u})=\alpha_s\left(\log|\mathcal{A}_s|+\sum_{a\in\mathcal{A}_s}u_a\log u_a\right)$ for any $s\in\mathcal{S}$, and (A), (B), (C) are satisfied. For any $s\in\mathcal{S}$ and $t=1,2,\cdots,T$, when $m_s^{(t)}$ is the trajectory Q-value feedback, we have

$$C_s^-=\frac{12\max_{s'\in\mathcal{S}}\left(\frac{2\|\boldsymbol{q}\|_\infty}{\alpha_{s'}}+\frac{\tau}{M_1}\log\frac{1}{\gamma}\right)}{\gamma}\qquad C_s^{/}=12\max_{s'\in\mathcal{S}}\left(\frac{2\|\boldsymbol{q}\|_\infty}{\alpha_{s'}}+\frac{\tau}{M_1}\log\frac{1}{\gamma}\right).\tag{C.2.5}$$

Proof.

$$\left|m_s^{(t+1)}-m_s^{(t)}\right|=\left|\frac{1}{\mu_1^{(t+1)}(\sigma(s))}-\frac{1}{\mu_1^{(t)}(\sigma(s))}\right|=\frac{1}{\mu_1^{(t+1)}(\sigma(s))}\left|\frac{\mu_1^{(t+1)}(\sigma(s))}{\mu_1^{(t)}(\sigma(s))}-1\right|.$$

We will then use the following lemma which shows the multiplicative stability when using negative entropy regularizer.

Lemma C.2.4. When $\psi_s^{\Delta}(\boldsymbol{u})=\alpha_s\left(\log|\mathcal{A}_s|+\sum_{a\in\mathcal{A}_s}u_a\log u_a\right)$ for any $s\in\mathcal{S}$, (A), (B), (C) are satisfied, then for any $s\in\mathcal{S},h\in s,t=1,2,\cdots,T$, we have

$$\left|\frac{\mu_1^{(t+1)}(\sigma_1(h))}{\mu_1^{(t)}(\sigma_1(h))}-1\right|,\left|\frac{\mu_2^{(t+1)}(\sigma_2(h))}{\mu_2^{(t)}(\sigma_2(h))}-1\right|\le 12\eta_s^{\mathrm{anc}}\max_{s'\in\mathcal{S}}\left(\frac{2\|\boldsymbol{q}\|_\infty}{\alpha_{s'}}+\frac{\tau}{M_1}\log\frac{1}{\gamma}\right).\tag{C.2.6}$$

The proof can be found at the end of this section. Then, for any $s\in\mathcal{S}_1$, we have

$$\begin{aligned}\left|m_s^{(t+1)}-m_s^{(t)}\right|=\frac{1}{\mu_1^{(t+1)}(\sigma(s))}\left|\frac{\mu_1^{(t+1)}(\sigma(s))}{\mu_1^{(t)}(\sigma(s))}-1\right|&\le\frac{12\eta_s^{\mathrm{anc}}\max_{s'\in\mathcal{S}}\left(\frac{2\|\boldsymbol{q}\|_\infty}{\alpha_{s'}}+\frac{\tau}{M_1}\log\frac{1}{\gamma}\right)}{\mu_1^{(t+1)}(\sigma(s))}\\&\le\frac{12\eta_s^{\mathrm{anc}}\max_{s'\in\mathcal{S}}\left(\frac{2\|\boldsymbol{q}\|_\infty}{\alpha_{s'}}+\frac{\tau}{M_1}\log\frac{1}{\gamma}\right)}{\gamma}.\end{aligned}$$

Therefore, $C_s^-=\frac{12\max_{s'\in\mathcal{S}}\left(\frac{2\|\boldsymbol{q}\|_\infty}{\alpha_{s'}}+\frac{\tau}{M_1}\log\frac{1}{\gamma}\right)}{\gamma}$.

Similarly, we have $C_s^{/}=12\max_{s'\in\mathcal{S}}\left(\frac{2\|\boldsymbol{q}\|_\infty}{\alpha_{s'}}+\frac{\tau}{M_1}\log\frac{1}{\gamma}\right)$. □

Proof of Lemma C.2.4. Firstly, we invoke Lemma C.2.5.

Lemma C.2.5. Consider update-rule (5.1.2). When $\psi_s^{\Delta}(\boldsymbol{u}) = \alpha_s \left(\log|\mathcal{A}| + \sum_{a\in\mathcal{A}_s} u_a \log u_a\right)$ for any $s \in \mathcal{S}$, and **(B)** is satisfied, for any $s \in \mathcal{S}$, $a \in \mathcal{A}_s$ and $t = 1, 2, \cdots, T$,

$$\exp\left(-\frac{\eta_s}{\alpha_s}\left(2\|\boldsymbol{q}\|_\infty + \frac{\tau\alpha_s}{M_1}\log\frac{1}{\gamma}\right)\right) \le \frac{\pi^{(t)}(a\,|\,s)}{\bar{\pi}^{(t)}(a\,|\,s)}, \frac{\bar{\pi}^{(t+1)}(a\,|\,s)}{\bar{\pi}^{(t)}(a\,|\,s)} \le \exp\left(\frac{\eta_s}{\alpha_s}\left(2\|\boldsymbol{q}\|_\infty + \frac{\tau\alpha_s}{M_1}\log\frac{1}{\gamma}\right)\right) \tag{C.2.7}$$

$$C_s^{\mathrm{diff}} = \frac{2}{\alpha_s}\left(2\|\boldsymbol{q}\|_\infty + \frac{\tau\alpha_s}{M_1}\log\frac{1}{\gamma}\right). \tag{C.2.8}$$

The proof is postponed to Appendix C.5.3.

By Lemma C.2.5, we have

$$\begin{aligned}
&\frac{\mu_1^{(t+1)}(\sigma_1(h))}{\mu_1^{(t)}(\sigma_1(h))} \\
=&\frac{\prod_{(s',a')\sqsubseteq\sigma_1(h)}\pi_1^{(t+1)}(a'\,|\,s')}{\prod_{(s',a')\sqsubseteq\sigma_1(h)}\pi_1^{(t)}(a'\,|\,s')} \\
=&\frac{\prod_{(s',a')\sqsubseteq\sigma_1(h)}\bar{\pi}_1^{(t+1)}(a'\,|\,s')}{\prod_{(s',a')\sqsubseteq\sigma_1(h)}\bar{\pi}_1^{(t)}(a'\,|\,s')} \cdot \frac{\prod_{(s',a')\sqsubseteq\sigma_1(h)}\pi_1^{(t+1)}(a'\,|\,s')}{\prod_{(s',a')\sqsubseteq\sigma_1(h)}\bar{\pi}_1^{(t+1)}(a'\,|\,s')} \cdot \frac{\prod_{(s',a')\sqsubseteq\sigma_1(h)}\bar{\pi}_1^{(t)}(a'\,|\,s')}{\prod_{(s',a')\sqsubseteq\sigma_1(h)}\pi_1^{(t)}(a'\,|\,s')} \\
\le&\exp\left(3\sum_{(s',a')\sqsubseteq\sigma_1(h)}\frac{\eta_{s'}}{\alpha_{s'}}\left(2\|\boldsymbol{q}\|_\infty + \frac{\tau\alpha_{s'}}{M_1}\log\frac{1}{\gamma}\right)\right) \\
\le&\exp\left(3\max_{s'\in\mathcal{S}}\left(\frac{2\|\boldsymbol{q}\|_\infty}{\alpha_{s'}} + \frac{\tau}{M_1}\log\frac{1}{\gamma}\right)\sum_{(s',a')\sqsubseteq\sigma_1(h)}\eta_{s'}\right) \\
\overset{\textbf{(A)}}{\le}&\exp\left(6\max_{s'\in\mathcal{S}}\left(\frac{2\|\boldsymbol{q}\|_\infty}{\alpha_{s'}} + \frac{\tau}{M_1}\log\frac{1}{\gamma}\right)\eta_s^{\mathrm{anc}}\right) \\
\le&1 + 12\eta_s^{\mathrm{anc}}\max_{s'\in\mathcal{S}}\left(\frac{2\|\boldsymbol{q}\|_\infty}{\alpha_{s'}} + \frac{\tau}{M_1}\log\frac{1}{\gamma}\right).
\end{aligned}$$

At the last line, we use the fact that $e^x \le 1 + 2x$ for $x \in [0, 1]$. Similarly, by using $1 + x \le e^x$, we can also get the lower-bound $1 - 6\eta_s^{\mathrm{anc}}\max_{s'\in\mathcal{S}}\left(\frac{2\|\boldsymbol{q}\|_\infty}{\alpha_{s'}} + \frac{\tau}{M_1}\log\frac{1}{\gamma}\right)$. □

C.2.2 Stability of Q-Value

Lemma C.2.6 (Stability of $m_s^{(t)}$ under Euclidean regularizer)**.** Consider when $\psi_s^{\Delta}(\boldsymbol{u}) = \frac{\alpha_s}{2}\sum_{a\in\mathcal{A}_s} u_a^2$ for any $s \in \mathcal{S}$, and **(A)** is satisfied. For any $s \in \mathcal{S}$ and $t = 1, 2, \cdots, T$, when $m_s^{(t)}$ is the Q-value feedback, we have

$$C_s^- = 6|s|\max_{s'\in\mathcal{S}} C_{s'}^{\mathrm{diff}} \qquad C_s^/ = \frac{6|s|}{M_1}\max_{s'\in\mathcal{S}} C_{s'}^{\mathrm{diff}} \tag{C.2.9}$$

where $|s|$ is the number of nodes in infoset s.

Proof. With Q-value feedback, for any $s \in \mathcal{S}_1$,

$$\begin{aligned}\left|m_s^{(t+1)} - m_s^{(t)}\right| =& \left|\sum_{h\in s} \mu_c(h)(\mu_2^{(t+1)}(\sigma_2(h)) - \mu_2^{(t)}(\sigma_2(h)))\right| \\ \le& \sum_{h\in s} \mu_c(h) \left|\mu_2^{(t+1)}(\sigma_2(h)) - \mu_2^{(t)}(\sigma_2(h))\right| \\ \le& |s| \max_{h\in s} \left|\mu_2^{(t+1)}(\sigma_2(h)) - \mu_2^{(t)}(\sigma_2(h))\right|.\end{aligned}$$

In the last line, $|s|$ denotes the number of nodes in s. By similar argument as in Lemma C.2.1, for any $h \in s$, we have

$$\begin{aligned}&\left|\mu_2^{(t+1)}(\sigma_2(h)) - \mu_2^{(t)}(\sigma_2(h))\right| \\ =& \left|\prod_{(s',a')\sqsubseteq\sigma_2(h)} \pi_2^{(t+1)}(a' \mid s') - \prod_{(s',a')\sqsubseteq\sigma_2(h)} \pi_2^{(t)}(a' \mid s')\right| \\ \le& \sum_{(s',a')\sqsubseteq\sigma_2(h)} \left|\pi_2^{(t+1)}(a' \mid s') - \pi_2^{(t)}(a' \mid s')\right| \\ \le& \sum_{(s',a')\sqsubseteq\sigma_2(h)} \Big(\left|\bar{\pi}_2^{(t+1)}(a' \mid s') - \bar{\pi}_2^{(t)}(a' \mid s')\right| + \left|\pi_2^{(t+1)}(a' \mid s') - \bar{\pi}_2^{(t+1)}(a' \mid s')\right| \\ &+ \left|\bar{\pi}_2^{(t)}(a' \mid s') - \pi_2^{(t)}(a' \mid s')\right| \Big) \\ \le& 3 \sum_{(s',a')\sqsubseteq\sigma_2(h)} C_{s'}^{\text{diff}} \eta_{s'}.\end{aligned}$$

Therefore, $\left|m_s^{(t+1)} - m_s^{(t)}\right| \le 3|s| \max_{h\in s} \sum_{(s',a')\sqsubseteq\sigma_2(h)} C_{s'}^{\text{diff}} \eta_{s'}$. Similarly,

$$\left|\frac{m_s^{(t+1)}}{m_s^{(t)}} - 1\right| = \frac{1}{m_s^{(t)}} \left|m_s^{(t+1)} - m_s^{(t)}\right| \le \frac{3|s|}{M_1} \max_{h\in s} \sum_{(s',a')\sqsubseteq\sigma_2(h)} C_{s'}^{\text{diff}} \eta_{s'}.$$

Finally, $C_s^- = 6|s| \max_{s'\in\mathcal{S}} C_{s'}^{\text{diff}}$ and $C_s' = \frac{6|s| \max_{s'\in\mathcal{S}} C_{s'}^{\text{diff}}}{M_1}$ according to **(A)**. □

Lemma C.2.7 (Stability of $m_s^{(t)}$ under Entropy regularizer)**.** Consider when $\psi_s^{\Delta}(\boldsymbol{u}) = \alpha_s \left(\log|\mathcal{A}_s| + \sum_{a\in\mathcal{A}_s} u_a \log u_a\right)$ for any $s \in \mathcal{S}$, and **(A)**, **(B)**, **(C)** are satisfied. For any $s \in \mathcal{S}$ and $t = 1, 2, \cdots, T$, when $m_s^{(t)}$ is the Q-value feedback, we have

$$C_s^- = 12M_2 \max_{s'\in\mathcal{S}} \left(\frac{2\|\boldsymbol{q}\|_\infty}{\alpha_{s'}} + \frac{\tau}{M_1} \log\frac{1}{\gamma}\right) \quad C_s' = 12 \max_{s'\in\mathcal{S}} \left(\frac{2\|\boldsymbol{q}\|_\infty}{\alpha_{s'}} + \frac{\tau}{M_1} \log\frac{1}{\gamma}\right). \quad \text{(C.2.10)}$$

Proof. With Q-value, for any $s \in \mathcal{S}_1$, we have

$$\left|\frac{m_s^{(t+1)}}{m_s^{(t)}} - 1\right| = \left|\frac{\sum_{h\in s} \mu_c(h)\mu_2^{(t+1)}(\sigma_2(h))}{\sum_{h\in s} \mu_c(h)\mu_2^{(t)}(\sigma_2(h))} - 1\right|.$$

By Lemma C.2.4, we have

$$\frac{\mu_2^{(t+1)}(\sigma_2(h))}{\mu_2^{(t)}(\sigma_2(h))} \le 1 + 12\eta_s^{\text{anc}} \max_{s' \in \mathcal{S}} \left(\frac{2\|\boldsymbol{q}\|_\infty}{\alpha_{s'}} + \frac{\tau}{M_1} \log \frac{1}{\gamma} \right)$$

Therefore,

$$\begin{aligned}
&\frac{\sum_{h \in s} \mu_c(h) \mu_2^{(t+1)}(\sigma_2(h))}{\sum_{h \in s} \mu_c(h) \mu_2^{(t)}(\sigma_2(h))} \\
\le &\frac{\sum_{h \in s} \mu_c(h) \left(1 + 12\eta_s^{\text{anc}} \max_{s' \in \mathcal{S}} \left(\frac{2\|\boldsymbol{q}\|_\infty}{\alpha_{s'}} + \frac{\tau}{M_1} \log \frac{1}{\gamma} \right)\right) \mu_2^{(t)}(\sigma_2(h))}{\sum_{h \in s} \mu_c(h) \mu_2^{(t)}(\sigma_2(h))} \\
\le &1 + 12\eta_s^{\text{anc}} \max_{s' \in \mathcal{S}} \left(\frac{2\|\boldsymbol{q}\|_\infty}{\alpha_{s'}} + \frac{\tau}{M_1} \log \frac{1}{\gamma} \right).
\end{aligned}$$

Similarly, we have

$$\frac{\sum_{h \in s} \mu_c(h) \mu_2^{(t+1)}(\sigma_2(h))}{\sum_{h \in s} \mu_c(h) \mu_2^{(t)}(\sigma_2(h))} \ge 1 - 12\eta_s^{\text{anc}} \max_{s' \in \mathcal{S}} \left(\frac{2\|\boldsymbol{q}\|_\infty}{\alpha_{s'}} + \frac{\tau}{M_1} \log \frac{1}{\gamma} \right).$$

Therefore, $\left| \frac{m_s^{(t+1)}}{m_s^{(t)}} - 1 \right| \le 12\eta_s^{\text{anc}} \max_{s' \in \mathcal{S}} \left(\frac{2\|\boldsymbol{q}\|_\infty}{\alpha_{s'}} + \frac{\tau}{M_1} \log \frac{1}{\gamma} \right)$. At the same time,

$$\left| m_s^{(t+1)} - m_s^{(t)} \right| = m_s^{(t)} \left| \frac{m_s^{(t+1)}}{m_s^{(t)}} - 1 \right| \le 12\eta_s^{\text{anc}} M_2 \max_{s' \in \mathcal{S}} \left(\frac{2\|\boldsymbol{q}\|_\infty}{\alpha_{s'}} + \frac{\tau}{M_1} \log \frac{1}{\gamma} \right). \qquad \square$$

C.3 Proof of Theorem 5.1.2

Theorem C.3.1 (Formal Version of Theorem 5.1.2)**.** Consider the update rule (5.1.2) and $q^{(t)}(s, \cdot)$ is chosen to be counterfactual value, trajectory Q-value, or Q-value. When $\frac{\eta_s^{\text{anc}}}{\eta_s} \le \tau C_s^\eta$, where $C_s^\eta := \frac{\gamma}{2C_s^-} \sum_{h \in s} \mu_c(h)$ for any $s \in \mathcal{S}$ and **(A)**, **(B)**, **(C)** are satisfied, we have the following guarantee.

$$\begin{aligned}
&\sum_{t=2}^{T} D_{\psi^\Pi}(\boldsymbol{\mu}^{(\tau,\gamma),*}, \boldsymbol{\mu}^{\overline{\boldsymbol{\pi}}^{(t)}}) \\
\le &\frac{2}{\gamma \min_{s \in \mathcal{S}} \sum_{h \in s} \mu_c(h)} \sum_{s \in \mathcal{S}} \left(C_s' + C_s^{-,Q} \right) \eta_s^{\text{anc}} \mu^{(\tau,\gamma),*}(\sigma(s)) \sum_{t=1}^{T} \left| \psi_s^\Delta(\pi_{p(s)}^{(t)}(\cdot \mid s)) - \psi_s^\Delta(\overline{\pi}_{p(s)}^{(t+1)}(\cdot \mid s)) \right| \\
&+ \frac{4}{\tau\gamma \min_{s \in \mathcal{S}} \sum_{h \in s} \mu_c(h)} \sum_{s \in \mathcal{S}} C_s^{\text{diff}} \mu^{(\tau,\gamma),*}(\sigma(s)) \|\boldsymbol{q}\|_\infty \eta_s M_2 T \qquad \text{(C.3.1)} \\
&+ \frac{2}{\tau\gamma \min_{s \in \mathcal{S}} \sum_{h \in s} \mu_c(h)} \sum_{s \in \mathcal{S}} \frac{m_s^{(1)}}{\eta_s} \mu^{(\tau,\gamma),*}(\sigma(s)) D_{\psi_s^\Delta}(\pi_{p(s)}^{(\tau,\gamma),*}(\cdot \mid s), \overline{\pi}_{p(s)}^{(1)}(\cdot \mid s)),
\end{aligned}$$

where $C_s^{-,Q}$ denotes C_s^- associated with Q-value, regardless of which feedback type $q^{(t)}(s, \cdot)$ is.

Proof Sketch. The structure of this section will be as follows. (i). By analyzing the update-rule (5.1.2), we can get the difference of utilities between our strategy $\boldsymbol{\pi}^{(t)}$ and an arbitrary strategy $\boldsymbol{\pi}$ at a single timestep t in each infoset. (ii). By telescoping and using the smoothness (the strategy as well as the feedback will not change much at each iteration) of the update-rule, we can further get an upperbound on the cumulated difference. (iii). By decomposition lemma [Liu et al., 2023], the difference in each infoset can be extended to the difference of utility in the whole game. Then, by rearranging the terms we can get an upperbound on the cumulated distance to the NE.

Firstly, we will use a standard analysis of the update rule (5.1.2). For notational simplicity, we define $\mu^{(t)}_{-p(s)}(s) \coloneqq \sum_{h\in s} \mu_c(h)\mu^{(t)}_{3-p(s)}(\sigma_{3-p(s)}(h))$.

Lemma C.3.2 (Generalized from Lemma C.2. in Liu et al. [2023])**.** Consider the update rule in (5.1.2). When ψ^{Δ}_s is strongly convex, then for any $\pi_{p(s)}(\cdot\,|\,s) \in \Delta^{|\mathcal{A}_s|}$ and $t \geq 1$, we have

$$
\begin{aligned}
&\eta_s \frac{\tau\mu^{(t)}_{-p(s)}(s)}{m^{(t)}_s}\psi^{\Delta}_s(\pi^{(t)}_{p(s)}(\cdot\,|\,s)) - \eta_s \frac{\tau\mu^{(t)}_{-p(s)}(s)}{m^{(t)}_s}\psi^{\Delta}_s(\pi_{p(s)}(\cdot\,|\,s)) \\
&+ \eta_s\tau\left(\frac{\mu^{(t-1)}_{-p(s)}(s)}{m^{(t-1)}_s} - \frac{\mu^{(t)}_{-p(s)}(s)}{m^{(t)}_s}\right)\left(\psi^{\Delta}_s(\pi^{(t)}_{p(s)}(\cdot\,|\,s)) - \psi^{\Delta}_s(\bar{\pi}^{(t+1)}_{p(s)}(\cdot\,|\,s))\right) \\
&+ \eta_s\left\langle -q^{(t)}(s,\cdot), \pi^{(t)}_{p(s)}(\cdot\,|\,s) - \pi_{p(s)}(\cdot\,|\,s)\right\rangle \\
\leq& D_{\psi^{\Delta}_s}(\pi_{p(s)}(\cdot\,|\,s), \bar{\pi}^{(t)}_{p(s)}(\cdot\,|\,s)) - (1+\eta_s\frac{\tau\mu^{(t)}_{-p(s)}(s)}{m^{(t)}_s})D_{\psi^{\Delta}_s}(\pi_{p(s)}(\cdot\,|\,s), \bar{\pi}^{(t+1)}_{p(s)}(\cdot\,|\,s)) \\
&- (1+\eta_s\frac{\tau\mu^{(t-1)}_{-p(s)}(s)}{m^{(t-1)}_s})D_{\psi^{\Delta}_s}(\bar{\pi}^{(t+1)}_{p(s)}(\cdot\,|\,s), \pi^{(t)}_{p(s)}(\cdot\,|\,s)) \\
&- D_{\psi^{\Delta}_s}(\pi^{(t)}_{p(s)}(\cdot\,|\,s), \bar{\pi}^{(t)}_{p(s)}(\cdot\,|\,s)) + \eta_s\left\langle q^{(t-1)}(s,\cdot) - q^{(t)}(s,\cdot), \pi^{(t)}_{p(s)}(\cdot\,|\,s) - \bar{\pi}^{(t+1)}_{p(s)}(\cdot\,|\,s)\right\rangle.
\end{aligned}
$$

The proof is postponed to Appendix C.3.2.

Multiplying $m^{(t)}_s$ on both sides of Lemma C.3.2, we have

$$
\begin{aligned}
&\eta_s\tau\mu^{(t)}_{-p(s)}(s)\psi^{\Delta}_s(\pi^{(t)}_{p(s)}(\cdot\,|\,s)) - \eta_s\tau\mu^{(t)}_{-p(s)}(s)\psi^{\Delta}_s(\pi_{p(s)}(\cdot\,|\,s)) \\
&+ \eta_s\tau\left(\frac{m^{(t)}_s}{m^{(t-1)}_s}\mu^{(t-1)}_{-p(s)}(s) - \mu^{(t)}_{-p(s)}(s)\right)\left(\psi^{\Delta}_s(\pi^{(t)}_{p(s)}(\cdot\,|\,s)) - \psi^{\Delta}_s(\bar{\pi}^{(t+1)}_{p(s)}(\cdot\,|\,s))\right) \\
&+ \eta_s m^{(t)}_s\langle -q^{(t)}(s,\cdot), \pi^{(t)}_{p(s)}(\cdot\,|\,s) - \pi_{p(s)}(\cdot\,|\,s)\rangle \\
\leq& m^{(t)}_s D_{\psi^{\Delta}_s}(\pi_{p(s)}(\cdot\,|\,s), \bar{\pi}^{(t)}_{p(s)}(\cdot\,|\,s)) - (m^{(t)}_s + \eta_s\tau\mu^{(t)}_{-p(s)}(s))D_{\psi^{\Delta}_s}(\pi_{p(s)}(\cdot\,|\,s), \bar{\pi}^{(t+1)}_{p(s)}(\cdot\,|\,s)) \\
&- (m^{(t)}_s + \eta_s\tau\frac{m^{(t)}_s}{m^{(t-1)}_s}\mu^{(t-1)}_{-p(s)}(s))D_{\psi^{\Delta}_s}(\bar{\pi}^{(t+1)}_{p(s)}(\cdot\,|\,s), \pi^{(t)}_{p(s)}(\cdot\,|\,s)) \\
&- m^{(t)}_s D_{\psi^{\Delta}_s}(\pi^{(t)}_{p(s)}(\cdot\,|\,s), \bar{\pi}^{(t)}_{p(s)}(\cdot\,|\,s)) + \eta_s m^{(t)}_s\left\langle q^{(t-1)}(s,\cdot) - q^{(t)}(s,\cdot), \pi^{(t)}_{p(s)}(\cdot\,|\,s) - \bar{\pi}^{(t+1)}_{p(s)}(\cdot\,|\,s)\right\rangle.
\end{aligned}
$$

By noticing the fact that $\mu^{(t)}_{-p(s)}(s)$ is equal to $m^{(t)}_s$ associated with Q-value, we can use Property 1 and Property 2 and get,

$$\begin{aligned}\left|\frac{m^{(t)}_s}{m^{(t-1)}_s}\mu^{(t-1)}_{-p(s)}(s)-\mu^{(t)}_{-p(s)}(s)\right| \leq& \left|\frac{m^{(t)}_s}{m^{(t-1)}_s}-1\right|\mu^{(t-1)}_{-p(s)}(s)+\left|\mu^{(t-1)}_{-p(s)}(s)-\mu^{(t)}_{-p(s)}(s)\right| \\ \leq& C^{/}_s\eta^{\text{anc}}_s+C^{-,Q}_s\eta^{\text{anc}}_s.\end{aligned}$$

We also use the fact that $\mu^{(t)}_{-p(s)}(s)\leq 1$ in the last inequality. We use $C^{-,Q}_s$ to denote the C^-_s associated with Q-value for simplicity.

Furthermore, by using Lemma C.2.2 and Hölder's Inequality, we have

$$\begin{aligned}&\left|\left\langle q^{(t-1)}(s,\cdot)-q^{(t)}(s,\cdot),\pi^{(t)}_{p(s)}(\cdot\,|\,s)-\overline{\pi}^{(t+1)}_{p(s)}(\cdot\,|\,s)\right\rangle\right| \\ \leq& \left\|q^{(t)}(s,\cdot)-q^{(t-1)}(s,\cdot)\right\|_\infty\cdot\left\|\pi^{(t)}_{p(s)}(\cdot\,|\,s)-\overline{\pi}^{(t+1)}_{p(s)}(\cdot\,|\,s)\right\|_1\leq 2C^{\text{diff}}_s\left\|\boldsymbol{q}\right\|_\infty\eta_s.\end{aligned}$$

where $\left\|\boldsymbol{q}\right\|_\infty=\max_{t\in[T],s\in\mathcal{S}}\left\|q^{(t)}(s,\cdot)\right\|_\infty$.

By telescoping and non-negativity of Bregman divergence, we have

$$\begin{aligned}&\sum_{t=1}^{T}\Big(\eta_s\tau\mu^{(t)}_{-p(s)}(s)\psi^\Delta_s(\pi^{(t)}_{p(s)}(\cdot\,|\,s))-\eta_s\tau\mu^{(t)}_{-p(s)}(s)\psi^\Delta_s(\pi_{p(s)}(\cdot\,|\,s)) \\ &+\eta_s m^{(t)}_s\langle -q^{(t)}(s,\cdot),\pi^{(t)}_{p(s)}(\cdot\,|\,s)-\pi_{p(s)}(\cdot\,|\,s)\rangle\Big) \\ \leq&\sum_{t=2}^{T}\underbrace{\left(m^{(t)}_s-m^{(t-1)}_s-\eta_s\tau\mu^{(t-1)}_{-p(s)}(s)\right)}_{①}D_{\psi^\Delta_s}(\pi_{p(s)}(\cdot\,|\,s),\overline{\pi}^{(t)}_{p(s)}(\cdot\,|\,s)) \\ &+\left(C^{/}_s+C^{-,Q}_s\right)\eta^{\text{anc}}_s\eta_s\tau\sum_{t=1}^{T}\left|\psi^\Delta_s(\pi^{(t)}_{p(s)}(\cdot\,|\,s))-\psi^\Delta_s(\overline{\pi}^{(t+1)}_{p(s)}(\cdot\,|\,s))\right|+2C^{\text{diff}}_s\left\|\boldsymbol{q}\right\|_\infty\eta^2_sM_2T \\ &+m^{(1)}_sD_{\psi^\Delta_s}(\pi_{p(s)}(\cdot\,|\,s),\overline{\pi}^{(1)}_{p(s)}(\cdot\,|\,s)).\end{aligned}$$

① can be upper-bounded by $C^-_s\eta^{\text{anc}}_s-\eta_s\tau\gamma\sum_{h\in s}\mu_c(h)\leq-\frac{\eta_s\tau\gamma}{2}\sum_{h\in s}\mu_c(h)$ by Property 2 and letting $\frac{\eta^{\text{anc}}_s}{\eta_s}\leq\frac{\tau\gamma}{2C^-_s}\sum_{h\in s}\mu_c(h)$. By non-negativity of Bregman divergence, we have

$$\begin{aligned}&\sum_{t=1}^{T}\Big(\eta_s\tau\mu^{(t)}_{-p(s)}(s)\psi^\Delta_s(\pi^{(t)}_{p(s)}(\cdot\,|\,s))-\eta_s\tau\mu^{(t)}_{-p(s)}(s)\psi^\Delta_s(\pi_{p(s)}(\cdot\,|\,s)) \\ &+\eta_s m^{(t)}_s\langle -q^{(t)}(s,\cdot),\pi^{(t)}_{p(s)}(\cdot\,|\,s)-\pi_{p(s)}(\cdot\,|\,s)\rangle\Big) \\ \leq&-\frac{\eta_s\tau\gamma\sum_{h\in s}\mu_c(h)}{2}\sum_{t=2}^{T}D_{\psi^\Delta_s}(\pi_{p(s)}(\cdot\,|\,s),\overline{\pi}^{(t)}_{p(s)}(\cdot\,|\,s)) \\ &+\left(C^{/}_s+C^{-,Q}_s\right)\eta_s\tau\eta^{\text{anc}}_s\sum_{t=1}^{T}\left|\psi^\Delta_s(\pi^{(t)}_{p(s)}(\cdot\,|\,s))-\psi^\Delta_s(\overline{\pi}^{(t+1)}_{p(s)}(\cdot\,|\,s))\right| \\ &+2C^{\text{diff}}_s\left\|\boldsymbol{q}\right\|_\infty\eta^2_sM_2T+m^{(1)}_sD_{\psi^\Delta_s}(\pi_{p(s)}(\cdot\,|\,s),\overline{\pi}^{(1)}_{p(s)}(\cdot\,|\,s)).\end{aligned}$$

Then, we will use the following regret decomposition lemma to extend the difference within an infoset above to the difference of the game.

Lemma C.3.3 (Lemma 5.1 in Liu et al. [2023])**.** Let $\Pi \coloneqq \Pi_1 \times \Pi_2$, the polytope of all valid sequence-form joint strategies. For any $\mu_1^{\pi_1} \in \Pi_1, \mu_2^{\pi_2} \in \Pi_2$, we let $\boldsymbol{\mu}^{\boldsymbol{\pi}} = (\mu_1^{\pi_1}, \mu_2^{\pi_2}) \in \Pi$ to denote the joint strategy, $\psi^{\Pi}(\boldsymbol{\mu}^{\boldsymbol{\pi}})\colon \Pi \to \mathbb{R} = \psi^{\Pi_1}(\mu_1^{\pi_1}) + \psi^{\Pi_2}(\mu_2^{\pi_2})$, and $F(\boldsymbol{\mu}^{\boldsymbol{\pi}}) \coloneqq (-\boldsymbol{A}\mu_2^{\pi_2}, \boldsymbol{A}^\top \mu_1^{\pi_1})$. For any $\boldsymbol{\mu}^{(1)}, \boldsymbol{\mu}^{(2)}, \cdots, \boldsymbol{\mu}^{(T)}, \boldsymbol{\mu}^{\boldsymbol{\pi}} \in \Pi$ and $\tau \geq 0$, we have

$$\begin{aligned} G^{(T),\Pi}(\boldsymbol{\mu}^{\boldsymbol{\pi}}) &\coloneqq \sum_{t=1}^{T} (F(\boldsymbol{\mu}^{(t)})^\top (\boldsymbol{\mu}^{(t)} - \boldsymbol{\mu}^{\boldsymbol{\pi}}) + \tau\psi^{\Pi}(\boldsymbol{\mu}^{(t)}) - \tau\psi^{\Pi}(\boldsymbol{\mu}^{\boldsymbol{\pi}})) \\ &= \sum_{s \in \mathcal{S}} \mu^{\boldsymbol{\pi}}(\sigma(s)) G^{(T)}(s; \pi_{p(s)}(\cdot \mid s)) \qquad \text{(C.3.2)} \\ R^{(T),\Pi} &\coloneqq \max_{\boldsymbol{\mu}^{\widehat{\boldsymbol{\pi}}} \in \Pi} G^{(T),\Pi}(\boldsymbol{\mu}^{\widehat{\boldsymbol{\pi}}}) \leq \max_{\boldsymbol{\mu}^{\widehat{\boldsymbol{\pi}}} \in \Pi} \sum_{s \in \mathcal{S}} \mu^{\widehat{\boldsymbol{\pi}}}(\sigma(s)) R^{(T)}(s) \qquad \text{(C.3.3)} \end{aligned}$$

where

$$\begin{aligned} B_p^{(t)}(s,a) &\coloneqq \sum_{(s,a) \sqsubseteq s'} \frac{\mu_p^{(t)}(\sigma(s'))}{\mu_p^{(t)}(s,a)} \psi_{s'}^{\Delta}(\pi_p^{(t)}(\cdot \mid s')) \qquad \text{(C.3.4)} \\ G^{(T)}(s; \pi_{p(s)}(\cdot \mid s)) &\coloneqq \sum_{t=1}^{T} \Big(\Big\langle -\mathsf{CF}_{p(s)}^{(t)}(s, \cdot) + \tau B_{p(s)}^{(t)}(s, \cdot), \pi_{p(s)}^{(t)}(\cdot \mid s) - \pi_{p(s)}(\cdot \mid s) \Big\rangle \\ &\qquad + \tau\psi_s^{\Delta}(\pi_{p(s)}^{(t)}(\cdot \mid s)) - \tau\psi_s^{\Delta}(\pi_{p(s)}(\cdot \mid s)) \Big) \qquad \text{(C.3.5)} \\ R^{(T)}(s) &\coloneqq \max_{\widehat{\pi}_{p(s)}(\cdot \mid s) \in \Delta^{|\mathcal{A}_s|}} G^{(T)}(s; \widehat{\pi}_{p(s)}(\cdot \mid s)). \qquad \text{(C.3.6)} \end{aligned}$$

By using Lemma C.3.3[1], we have

$$
\begin{aligned}
&\sum_{s\in\mathcal{S}}\mu^{(\tau,\gamma),*}(\sigma(s))\sum_{t=1}^{T}\Big(\tau\mu^{(t)}_{-p(s)}(s)\psi^{\Delta}_{s}(\pi^{(t)}_{p(s)}(\cdot\,|\,s))-\tau\mu^{(t)}_{-p(s)}(s)\psi^{\Delta}_{s}(\pi^{(\tau,\gamma),*}_{p(s)}(\cdot\,|\,s))\\
&\quad+m^{(t)}_{s}\langle -q^{(t)}(s,\cdot),\pi^{(t)}_{p(s)}(\cdot\,|\,s)-\pi^{(\tau,\gamma),*}_{p(s)}(\cdot\,|\,s)\rangle\Big)\\
=&\sum_{t=1}^{T}\Big(\left(\mu_1^{(\tau,\gamma),*}-\mu_1^{(t)}\right)^{\top}\boldsymbol{A}\mu_2^{(t)}+\left(\mu_1^{(t)}\right)^{\top}\boldsymbol{A}\left(\mu_2^{(t)}-\mu_2^{(\tau,\gamma),*}\right)\\
&\quad+\tau\left(\psi^{\Pi_1}_{\mathrm{bi}}(\mu_1^{(t)},\mu_2^{(t)})-\psi^{\Pi_1}_{\mathrm{bi}}(\mu_1^{(\tau,\gamma),*},\mu_2^{(t)})-\psi^{\Pi_2}_{\mathrm{bi}}(\mu_1^{(t)},\mu_2^{(t)})+\psi^{\Pi_2}_{\mathrm{bi}}(\mu_1^{(\tau,\gamma),*},\mu_2^{(t)})\right)\\
&\quad+\tau\left(\psi^{\Pi_2}_{\mathrm{bi}}(\mu_1^{(t)},\mu_2^{(t)})-\psi^{\Pi_2}_{\mathrm{bi}}(\mu_1^{(t)},\mu_2^{(\tau,\gamma),*})-\psi^{\Pi_1}_{\mathrm{bi}}(\mu_1^{(t)},\mu_2^{(t)})+\psi^{\Pi_1}_{\mathrm{bi}}(\mu_1^{(t)},\mu_2^{(\tau,\gamma),*})\right)\Big)\\
=&\sum_{t=1}^{T}\Big(\left(\mu_1^{(\tau,\gamma),*}-\mu_1^{(t)}\right)^{\top}\boldsymbol{A}\mu_2^{(\tau,\gamma),*}+\left(\mu_1^{(\tau,\gamma),*}\right)^{\top}\boldsymbol{A}\left(\mu_2^{(t)}-\mu_2^{(\tau,\gamma),*}\right)\\
&\quad+\tau\left(\psi^{\Pi_1}_{\mathrm{bi}}(\mu_1^{(t)},\mu_2^{(\tau,\gamma),*})-\psi^{\Pi_1}_{\mathrm{bi}}(\mu_1^{(\tau,\gamma),*},\mu_1^{(\tau,\gamma),*})+\psi^{\Pi_2}_{\mathrm{bi}}(\mu_1^{(\tau,\gamma),*},\mu_1^{(\tau,\gamma),*})-\psi^{\Pi_2}_{\mathrm{bi}}(\mu_1^{(t)},\mu_2^{(\tau,\gamma),*})\right)\\
&\quad+\tau\left(\psi^{\Pi_2}_{\mathrm{bi}}(\mu_1^{(\tau,\gamma),*},\mu_2^{(t)})-\psi^{\Pi_2}_{\mathrm{bi}}(\mu_1^{(\tau,\gamma),*},\mu_2^{(\tau,\gamma),*})+\psi^{\Pi_1}_{\mathrm{bi}}(\mu_1^{(\tau,\gamma),*},\mu_1^{(\tau,\gamma),*})-\psi^{\Pi_1}_{\mathrm{bi}}(\mu_1^{(\tau,\gamma),*},\mu_2^{(t)})\right)\Big)\\
\geq&0.
\end{aligned}
$$

The last inequality is because $\boldsymbol{\mu}^{(\tau,\gamma),*}$ is the NE of $\max_{\mu_1^{\pi_1}\in\Pi_1\colon \mu_1^{\pi_1}\succeq\gamma}\min_{\mu_2^{\pi_2}\in\Pi_2\colon \mu_2^{\pi_2}\succeq\gamma}\left(\mu_1^{\pi_1}\right)^{\top}\boldsymbol{A}\mu_2^{\pi_2}-\tau\psi^{\Pi_1}_{\mathrm{bi}}(\mu_1^{\pi_1},\mu_2^{\pi_2})+\tau\psi^{\Pi_2}_{\mathrm{bi}}(\mu_1^{\pi_1},\mu_2^{\pi_2})$.

Therefore,

$$
\begin{aligned}
0\leq&\sum_{s\in\mathcal{S}}\mu^{(\tau,\gamma),*}(\sigma(s))\sum_{t=1}^{T}\Big(\tau\mu^{(t)}_{-p(s)}(s)\psi^{\Delta}_{s}(\pi^{(t)}_{p(s)}(\cdot\,|\,s))-\tau\mu^{(t)}_{-p(s)}(s)\psi^{\Delta}_{s}(\pi^{(\tau,\gamma),*}_{p(s)}(\cdot\,|\,s))\\
&\quad+m^{(t)}_{s}\langle -q^{(t)}(s,\cdot),\pi^{(t)}_{p(s)}(\cdot\,|\,s)-\pi^{(\tau,\gamma),*}_{p(s)}(\cdot\,|\,s)\rangle\Big)\\
\leq&-\frac{\tau\gamma\min_{s\in\mathcal{S}}\sum_{h\in s}\mu_c(h)}{2}\sum_{t=2}^{T}\sum_{s\in\mathcal{S}}\mu^{(\tau,\gamma),*}(\sigma(s))D_{\psi^{\Delta}_{s}}(\pi^{(\tau,\gamma),*}_{p(s)}(\cdot\,|\,s),\overline{\pi}^{(t)}_{p(s)}(\cdot\,|\,s))\\
&\quad+\sum_{s\in\mathcal{S}}\left(C^{\prime}_{s}+C^{-,Q}_{s}\right)\tau\eta^{\mathrm{anc}}_{s}\mu^{(\tau,\gamma),*}(\sigma(s))\sum_{t=1}^{T}\left|\psi^{\Delta}_{s}(\pi^{(t)}_{p(s)}(\cdot\,|\,s))-\psi^{\Delta}_{s}(\overline{\pi}^{(t+1)}_{p(s)}(\cdot\,|\,s))\right|\\
&\quad+2\sum_{s\in\mathcal{S}}C^{\mathrm{diff}}_{s}\mu^{(\tau,\gamma),*}(\sigma(s))\left\|\boldsymbol{q}\right\|_{\infty}\eta_s M_2 T\\
&\quad+\sum_{s\in\mathcal{S}}\frac{m^{(1)}_{s}}{\eta_s}\mu^{(\tau,\gamma),*}(\sigma(s))D_{\psi^{\Delta}_{s}}(\pi^{(\tau,\gamma),*}_{p(s)}(\cdot\,|\,s),\overline{\pi}^{(1)}_{p(s)}(\cdot\,|\,s)).
\end{aligned}
$$

(i) is because $\boldsymbol{\mu}^{(\tau,\gamma),*}$ is the NE of the regularized and perturbed EFG. Then, by rearranging

[1]It can be easily generalized to bidilated version by absorbing the reach probability of player $3-p$ and the chance player into ψ^{Δ}_{s} for $\psi^{\Pi_p}_{\mathrm{bi}}$ and player p. For $\psi^{\Pi_{3-p}}_{\mathrm{bi}}$, since it is linear with respect to $\mu_p^{\pi_p}$, it can be combined into the counterfactual value.

the terms, we have

$$
\begin{aligned}
&\sum_{t=2}^{T} D_{\psi^{\Pi}}(\boldsymbol{\mu}^{(\tau,\gamma),*}, \boldsymbol{\mu}^{\overline{\boldsymbol{\pi}}^{(t)}}) \\
=&\sum_{t=2}^{T}\sum_{s\in\mathcal{S}} \mu^{(\tau,\gamma),*}(\sigma(s)) D_{\psi_s^{\Delta}}(\pi_{p(s)}^{(\tau,\gamma),*}(\cdot\,|\,s), \overline{\pi}_{p(s)}^{(t)}(\cdot\,|\,s)) \\
\leq&\frac{2}{\gamma \min_{s\in\mathcal{S}} \sum_{h\in s}\mu_c(h)} \sum_{s\in\mathcal{S}} \left(C_s' + C_s^{-,Q}\right) \eta_s^{\mathrm{anc}} \mu^{(\tau,\gamma),*}(\sigma(s)) \sum_{t=1}^{T} \left|\psi_s^{\Delta}(\pi_{p(s)}^{(t)}(\cdot\,|\,s)) - \psi_s^{\Delta}(\overline{\pi}_{p(s)}^{(t+1)}(\cdot\,|\,s))\right| \\
&+ \frac{4}{\tau\gamma \min_{s\in\mathcal{S}} \sum_{h\in s}\mu_c(h)} \sum_{s\in\mathcal{S}} C_s^{\mathrm{diff}} \mu^{(\tau,\gamma),*}(\sigma(s)) \left\|\boldsymbol{q}\right\|_{\infty} \eta_s M_2 T \\
&+ \frac{2}{\tau\gamma \min_{s\in\mathcal{S}} \sum_{h\in s}\mu_c(h)} \sum_{s\in\mathcal{S}} \frac{m_s^{(1)}}{\eta_s} \mu^{(\tau,\gamma),*}(\sigma(s)) D_{\psi_s^{\Delta}}(\pi_{p(s)}^{(\tau,\gamma),*}(\cdot\,|\,s), \overline{\pi}_{p(s)}^{(1)}(\cdot\,|\,s)).
\end{aligned}
$$

The first equality is by Lemma 5.1.1. Now, we achieved best-iterate convergence to the regularized NE $\boldsymbol{\mu}^{(\tau,\gamma),*}$ in terms of Bregman divergence. □

C.3.1 Proof of Lemma 5.1.1

By definition of Bregman divergence, we have

$$
D_{\psi^{\Pi}}(\boldsymbol{\mu}^{\boldsymbol{\pi}}, \boldsymbol{\mu}^{\widetilde{\boldsymbol{\pi}}}) = \psi^{\Pi}(\boldsymbol{\mu}^{\boldsymbol{\pi}}) - \psi^{\Pi}(\boldsymbol{\mu}^{\widetilde{\boldsymbol{\pi}}}) - \left\langle \nabla\psi^{\Pi}(\boldsymbol{\mu}^{\widetilde{\boldsymbol{\pi}}}), \boldsymbol{\mu}^{\boldsymbol{\pi}} - \boldsymbol{\mu}^{\widetilde{\boldsymbol{\pi}}} \right\rangle.
$$

For notational simplicity, we use $\mu^{\boldsymbol{\pi}}(s,a)$ as $\mu_{p(s)}^{\boldsymbol{\pi}}(s,a)$. Since $\psi^{\Pi}(\boldsymbol{\mu}^{\widetilde{\boldsymbol{\pi}}}) = \sum_{s\in\mathcal{S}} \mu^{\widetilde{\boldsymbol{\pi}}}(\sigma(s)) \psi_s^{\Delta}\left(\frac{\mu^{\widetilde{\boldsymbol{\pi}}}(s,\cdot)}{\mu^{\widetilde{\boldsymbol{\pi}}}(\sigma(s))}\right)$, for any $s\in\mathcal{S}, a\in\mathcal{A}_s$, the gradient $\nabla_{\mu^{\widetilde{\boldsymbol{\pi}}}(s,a)}\psi^{\Pi}(\boldsymbol{\mu}^{\widetilde{\boldsymbol{\pi}}})$ is equal to

$$
\begin{aligned}
\nabla_{\mu^{\widetilde{\boldsymbol{\pi}}}(s,a)}\psi^{\Pi}(\boldsymbol{\mu}^{\widetilde{\boldsymbol{\pi}}}) = &\sum_{s'\in\mathcal{S}:\ \sigma(s')=(s,a)} \left(\psi_{s'}^{\Delta}\left(\frac{\mu^{\widetilde{\boldsymbol{\pi}}}(s',\cdot)}{\mu^{\widetilde{\boldsymbol{\pi}}}(\sigma(s'))}\right) - \left\langle \nabla\psi_{s'}^{\Delta}\left(\frac{\mu^{\widetilde{\boldsymbol{\pi}}}(s',\cdot)}{\mu^{\widetilde{\boldsymbol{\pi}}}(\sigma(s'))}\right), \frac{\mu^{\widetilde{\boldsymbol{\pi}}}(s',\cdot)}{\mu^{\widetilde{\boldsymbol{\pi}}}(\sigma(s'))} \right\rangle\right) \\
&+ \nabla_a \psi_s^{\Delta}\left(\frac{\mu^{\widetilde{\boldsymbol{\pi}}}(s,\cdot)}{\mu^{\widetilde{\boldsymbol{\pi}}}(\sigma(s))}\right).
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
&\left\langle \nabla \psi^{\Pi}(\boldsymbol{\mu}^{\widetilde{\boldsymbol{\pi}}}), \boldsymbol{\mu}^{\widetilde{\boldsymbol{\pi}}} \right\rangle \\
&= \sum_{s\in\mathcal{S}, a\in\mathcal{A}_s} \mu^{\widetilde{\boldsymbol{\pi}}}(s,a) \sum_{s'\in\mathcal{S}\colon \sigma(s')=(s,a)} \left(\psi_{s'}^{\Delta}\left(\frac{\mu^{\widetilde{\boldsymbol{\pi}}}(s',\cdot)}{\mu^{\widetilde{\boldsymbol{\pi}}}(\sigma(s'))} \right) - \left\langle \nabla \psi_{s'}^{\Delta}\left(\frac{\mu^{\widetilde{\boldsymbol{\pi}}}(s',\cdot)}{\mu^{\widetilde{\boldsymbol{\pi}}}(\sigma(s'))} \right), \frac{\mu^{\widetilde{\boldsymbol{\pi}}}(s',\cdot)}{\mu^{\widetilde{\boldsymbol{\pi}}}(\sigma(s'))} \right\rangle \right) \\
&\quad + \sum_{s\in\mathcal{S}, a\in\mathcal{A}_s} \mu^{\widetilde{\boldsymbol{\pi}}}(s,a) \nabla_a \psi_s^{\Delta}\left(\frac{\mu^{\widetilde{\boldsymbol{\pi}}}(s,\cdot)}{\mu^{\widetilde{\boldsymbol{\pi}}}(\sigma(s))} \right) \\
&= \sum_{s\in\mathcal{S}, a\in\mathcal{A}_s} \sum_{s'\in\mathcal{S}\colon \sigma(s')=(s,a)} \mu^{\widetilde{\boldsymbol{\pi}}}(\sigma(s')) \psi_{s'}^{\Delta}\left(\frac{\mu^{\widetilde{\boldsymbol{\pi}}}(s',\cdot)}{\mu^{\widetilde{\boldsymbol{\pi}}}(\sigma(s'))} \right) && \text{(C.3.7)} \\
&\quad - \sum_{s\in\mathcal{S}, a\in\mathcal{A}_s} \sum_{s'\in\mathcal{S}\colon \sigma(s')=(s,a)} \left\langle \nabla \psi_{s'}^{\Delta}\left(\frac{\mu^{\widetilde{\boldsymbol{\pi}}}(s',\cdot)}{\mu^{\widetilde{\boldsymbol{\pi}}}(\sigma(s'))} \right), \mu^{\widetilde{\boldsymbol{\pi}}}(s',\cdot) \right\rangle && \text{(C.3.8)} \\
&\quad + \sum_{s\in\mathcal{S}, a\in\mathcal{A}_s} \mu^{\widetilde{\boldsymbol{\pi}}}(s,a) \nabla_a \psi_s^{\Delta}\left(\frac{\mu^{\widetilde{\boldsymbol{\pi}}}(s,\cdot)}{\mu^{\widetilde{\boldsymbol{\pi}}}(\sigma(s))} \right). && \text{(C.3.9)}
\end{aligned}
$$

Note that (C.3.7) is equal to $\sum_{s\in\mathcal{S}} \mu^{\widetilde{\boldsymbol{\pi}}}(\sigma(s)) \psi_s^{\Delta}\left(\frac{\mu^{\widetilde{\boldsymbol{\pi}}}(s,\cdot)}{\mu^{\widetilde{\boldsymbol{\pi}}}(\sigma(s))} \right) = \psi^{\Pi}(\boldsymbol{\mu}^{\widetilde{\boldsymbol{\pi}}})$ due to the uniqueness of $\sigma(s')$.

Similarly, due to the uniqueness of $\sigma(s')$, (C.3.8) is equal to $-\sum_{s\in\mathcal{S}} \left\langle \nabla \psi_s^{\Delta}\left(\frac{\mu^{\widetilde{\boldsymbol{\pi}}}(s,\cdot)}{\mu^{\widetilde{\boldsymbol{\pi}}}(\sigma(s))} \right), \mu^{\widetilde{\boldsymbol{\pi}}}(s,\cdot) \right\rangle$, which is equal to the negative of (C.3.9) and thus cancel out. Therefore,

$$
\left\langle \nabla \psi^{\Pi}(\boldsymbol{\mu}^{\widetilde{\boldsymbol{\pi}}}), \boldsymbol{\mu}^{\widetilde{\boldsymbol{\pi}}} \right\rangle = \psi^{\Pi}(\boldsymbol{\mu}^{\widetilde{\boldsymbol{\pi}}}).
$$

Moreover,

$$
\begin{aligned}
&\left\langle \nabla \psi^{\Pi}(\boldsymbol{\mu}^{\widetilde{\boldsymbol{\pi}}}), \boldsymbol{\mu}^{\boldsymbol{\pi}} \right\rangle \\
&= \sum_{s\in\mathcal{S}, a\in\mathcal{A}_s} \sum_{s'\in\mathcal{S}\colon \sigma(s')=(s,a)} \mu^{\boldsymbol{\pi}}(\sigma(s')) \psi_{s'}^{\Delta}\left(\frac{\mu^{\widetilde{\boldsymbol{\pi}}}(s',\cdot)}{\mu^{\widetilde{\boldsymbol{\pi}}}(\sigma(s'))} \right) \\
&\quad - \sum_{s\in\mathcal{S}, a\in\mathcal{A}_s} \mu^{\boldsymbol{\pi}}(s,a) \sum_{s'\in\mathcal{S}\colon \sigma(s')=(s,a)} \left\langle \nabla \psi_{s'}^{\Delta}\left(\frac{\mu^{\widetilde{\boldsymbol{\pi}}}(s',\cdot)}{\mu^{\widetilde{\boldsymbol{\pi}}}(\sigma(s'))} \right), \frac{\mu^{\widetilde{\boldsymbol{\pi}}}(s',\cdot)}{\mu^{\widetilde{\boldsymbol{\pi}}}(\sigma(s'))} \right\rangle \\
&\quad + \sum_{s\in\mathcal{S}, a\in\mathcal{A}_s} \mu^{\boldsymbol{\pi}}(s,a) \nabla_a \psi_s^{\Delta}\left(\frac{\mu^{\widetilde{\boldsymbol{\pi}}}(s,\cdot)}{\mu^{\widetilde{\boldsymbol{\pi}}}(s,a)} \right) \\
&= \sum_{s\in\mathcal{S}} \mu^{\boldsymbol{\pi}}(\sigma(s)) \left(\psi_s^{\Delta}\left(\frac{\mu^{\widetilde{\boldsymbol{\pi}}}(s,\cdot)}{\mu^{\widetilde{\boldsymbol{\pi}}}(\sigma(s))} \right) + \left\langle \nabla \psi_s^{\Delta}\left(\frac{\mu^{\widetilde{\boldsymbol{\pi}}}(s,\cdot)}{\mu^{\widetilde{\boldsymbol{\pi}}}(\sigma(s))} \right), \frac{\mu^{\boldsymbol{\pi}}(s,\cdot)}{\mu^{\boldsymbol{\pi}}(\sigma(s))} - \frac{\mu^{\widetilde{\boldsymbol{\pi}}}(s,\cdot)}{\mu^{\widetilde{\boldsymbol{\pi}}}(\sigma(s))} \right\rangle \right).
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
&\left\langle F^{\tau}(\boldsymbol{\mu}^{\boldsymbol{\pi}}), \boldsymbol{\mu}^{\boldsymbol{\pi}} - \boldsymbol{\mu}^{\boldsymbol{\pi}'} \right\rangle \\
&= -(\mu_1^{\pi_1})^{\top} \boldsymbol{A} \mu_2^{\pi_2'} + (\mu_1^{\pi_1'})^{\top} \boldsymbol{A} \mu_2^{\pi_2} + \tau D_{\psi_{\mathrm{bi}}^{\Pi_1}(\cdot,\mu_2^{\pi_2})}\left(\mu_1^{\pi_1'}, \mu_1^{\pi_1} \right) + \tau D_{\psi_{\mathrm{bi}}^{\Pi_2}(\mu_1^{\pi_1},\cdot)}\left(\mu_2^{\pi_2'}, \mu_2^{\pi_2} \right) \\
&\quad + \tau \psi_{\mathrm{bi}}^{\Pi_1}(\mu_1^{\pi_1}, \mu_2^{\pi_2'}) - \tau \psi_{\mathrm{bi}}^{\Pi_1}(\mu_1^{\pi_1'}, \mu_2^{\pi_2}) + \tau \psi_{\mathrm{bi}}^{\Pi_2}(\mu_1^{\pi_1'}, \mu_2^{\pi_2}) - \tau \psi_{\mathrm{bi}}^{\Pi_2}(\mu_1^{\pi_1}, \mu_2^{\pi_2'}).
\end{aligned}
$$

C.3.2 Proof of Lemma C.3.2

Firstly, we introduce the following lemma.

Lemma C.3.4. Let $\mathcal{C}$ be a convex set and $\boldsymbol{x}^{(1)} = \operatorname{argmin}_{\boldsymbol{x}\in\mathcal{C}} \left\{ \langle \boldsymbol{g}, \boldsymbol{x} \rangle + \tau_0 \psi^{\mathcal{C}}(\boldsymbol{x}) + \frac{1}{\eta} D_{\psi^{\mathcal{C}}}(\boldsymbol{x}, \boldsymbol{x}^{(0)}) \right\}$, where $\psi^{\mathcal{C}}$ is a strongly-convex function in $\mathcal{C}$ and $\tau_0 \geq 0$ is a constant. Then, for any $\boldsymbol{x}^{(2)} \in \mathcal{C}$, we have

$$\begin{aligned} &\eta\tau_0\psi^{\mathcal{C}}(\boldsymbol{x}^{(1)}) - \eta\tau_0\psi^{\mathcal{C}}(\boldsymbol{x}^{(2)}) + \eta \left\langle \boldsymbol{g}, \boldsymbol{x}^{(1)} - \boldsymbol{x}^{(2)} \right\rangle \\ \leq & D_{\psi^{\mathcal{C}}}(\boldsymbol{x}^{(2)}, \boldsymbol{x}^{(0)}) - (1+\eta\tau_0) D_{\psi^{\mathcal{C}}}(\boldsymbol{x}^{(2)}, \boldsymbol{x}^{(1)}) - D_{\psi^{\mathcal{C}}}(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(0)}). \end{aligned} \tag{C.3.10}$$

The proof is postponed to the end of this section.

Plug $\boldsymbol{x}^{(0)} = \bar{\pi}^{(t)}_{p(s)}(\cdot \,|\, s), \boldsymbol{x}^{(1)} = \bar{\pi}^{(t+1)}_{p(s)}(\cdot \,|\, s), \boldsymbol{x}^{(2)} = \pi_{p(s)}(\cdot \,|\, s), \boldsymbol{g} = -q^{(t)}(s,\cdot), \psi^{\mathcal{C}} = \psi^{\Delta}_s, \eta = \eta_s, \tau_0 = \frac{\tau \mu^{(t)}_{-p(s)}(s)}{m^{(t)}_s}$ into Lemma C.3.4, with $\mathcal{C} = \Delta^{|\mathcal{A}_s|}_{\gamma_s, \boldsymbol{\nu}_s}$,

$$\begin{aligned} &\eta_s \frac{\tau \mu^{(t)}_{-p(s)}(s)}{m^{(t)}_s} \psi^{\Delta}_s(\bar{\pi}^{(t+1)}_{p(s)}(\cdot \,|\, s)) - \eta_s \frac{\tau \mu^{(t)}_{-p(s)}(s)}{m^{(t)}_s} \psi^{\Delta}_s(\pi_{p(s)}(\cdot \,|\, s)) \\ &+ \eta_s \langle \bar{\pi}^{(t+1)}_{p(s)}(\cdot \,|\, s) - \pi_{p(s)}(\cdot \,|\, s), -q^{(t)}(s,\cdot) \rangle \\ \leq & D_{\psi^{\Delta}_s}(\pi_{p(s)}(\cdot \,|\, s), \bar{\pi}^{(t)}_{p(s)}(\cdot \,|\, s)) - \left(1 + \eta_s \frac{\tau \mu^{(t)}_{-p(s)}(s)}{m^{(t)}_s}\right) D_{\psi^{\Delta}_s}(\pi_{p(s)}(\cdot \,|\, s), \bar{\pi}^{(t+1)}_{p(s)}(\cdot \,|\, s)) \\ &- D_{\psi^{\Delta}_s}(\bar{\pi}^{(t+1)}_{p(s)}(\cdot \,|\, s), \bar{\pi}^{(t)}_{p(s)}(\cdot \,|\, s)). \end{aligned}$$

Plug $\boldsymbol{x}^{(0)} = \bar{\pi}^{(t)}_{p(s)}(\cdot \,|\, s), \boldsymbol{x}^{(1)} = \pi^{(t)}_{p(s)}(\cdot \,|\, s), \boldsymbol{x}^{(2)} = \bar{\pi}^{(t+1)}_{p(s)}(\cdot \,|\, s), \boldsymbol{g} = -q^{(t-1)}(s,\cdot), \psi^{\mathcal{C}} = \psi^{\Delta}_s, \eta = \eta_s, \tau_0 = \frac{\tau \mu^{(t-1)}_{-p(s)}(s)}{m^{(t-1)}_s}$ into Lemma C.3.4, with $\mathcal{C} = \Delta^{|\mathcal{A}_s|}_{\gamma_s, \boldsymbol{\nu}_s}$,

$$\begin{aligned} &\eta_s \frac{\tau \mu^{(t-1)}_{-p(s)}(s)}{m^{(t-1)}_s} \psi^{\Delta}_s(\pi^{(t)}_{p(s)}(\cdot \,|\, s)) - \eta_s \frac{\tau \mu^{(t-1)}_{-p(s)}(s)}{m^{(t-1)}_s} \psi^{\Delta}_s(\bar{\pi}^{(t+1)}_{p(s)}(\cdot \,|\, s)) \\ &+ \eta_s \langle \pi^{(t)}_{p(s)}(\cdot \,|\, s) - \bar{\pi}^{(t+1)}_{p(s)}(\cdot \,|\, s), -q^{(t-1)}(s,\cdot) \rangle \\ \leq & D_{\psi^{\Delta}_s}(\bar{\pi}^{(t+1)}_{p(s)}(\cdot \,|\, s), \bar{\pi}^{(t)}_{p(s)}(\cdot \,|\, s)) - (1 + \eta_s \frac{\tau \mu^{(t-1)}_{-p(s)}(s)}{m^{(t-1)}_s}) D_{\psi^{\Delta}_s}(\bar{\pi}^{(t+1)}_{p(s)}(\cdot \,|\, s), \pi^{(t)}_{p(s)}(\cdot \,|\, s)) \\ &- D_{\psi^{\Delta}_s}(\pi^{(t)}_{p(s)}(\cdot \,|\, s), \bar{\pi}^{(t)}_{p(s)}(\cdot \,|\, s)). \end{aligned}$$

Summing them up and adding $\eta_s \left\langle q^{(t-1)}(s,\cdot) - q^{(t)}(s,\cdot), \pi^{(t)}_{p(s)}(\cdot \,|\, s) - \bar{\pi}^{(t+1)}_{p(s)}(\cdot \,|\, s) \right\rangle$ to both

sides, we have

$$
\begin{aligned}
&\eta_s \frac{\tau \mu^{(t)}_{-p(s)}(s)}{m^{(t)}_s} \psi^{\Delta}_s(\pi^{(t)}_{p(s)}(\cdot \,|\, s)) - \eta_s \frac{\tau \mu^{(t)}_{-p(s)}(s)}{m^{(t)}_s} \psi^{\Delta}_s(\pi_{p(s)}(\cdot \,|\, s)) \qquad \text{(C.3.11)}\\
&+ \eta_s \tau \Big(\frac{\mu^{(t-1)}_{-p(s)}(s)}{m^{(t-1)}_s} - \frac{\mu^{(t)}_{-p(s)}(s)}{m^{(t)}_s} \Big) \big(\psi^{\Delta}_s(\pi^{(t)}_{p(s)}(\cdot \,|\, s)) - \psi^{\Delta}_s(\overline{\pi}^{(t+1)}_{p(s)}(\cdot \,|\, s)) \big)\\
&+ \eta_s \langle -q^{(t)}(s, \cdot), \pi^{(t)}_{p(s)}(\cdot \,|\, s) - \pi_{p(s)}(\cdot \,|\, s) \rangle\\
\leq & D_{\psi^{\Delta}_s}(\pi_{p(s)}(\cdot \,|\, s), \overline{\pi}^{(t)}_{p(s)}(\cdot \,|\, s)) - \Big(1 + \eta_s \frac{\tau \mu^{(t)}_{-p(s)}(s)}{m^{(t)}_s}\Big) D_{\psi^{\Delta}_s}(\pi_{p(s)}(\cdot \,|\, s), \overline{\pi}^{(t+1)}_{p(s)}(\cdot \,|\, s))\\
&- \Big(1 + \eta_s \frac{\tau \mu^{(t-1)}_{-p(s)}(s)}{m^{(t-1)}_s}\Big) D_{\psi^{\Delta}_s}(\overline{\pi}^{(t+1)}_{p(s)}(\cdot \,|\, s), \pi^{(t)}_{p(s)}(\cdot \,|\, s))\\
&- D_{\psi^{\Delta}_s}(\pi^{(t)}_{p(s)}(\cdot \,|\, s), \overline{\pi}^{(t)}_{p(s)}(\cdot \,|\, s)) + \eta_s \left\langle q^{(t-1)}(s, \cdot) - q^{(t)}(s, \cdot), \pi^{(t)}_{p(s)}(\cdot \,|\, s) - \overline{\pi}^{(t+1)}_{p(s)}(\cdot \,|\, s) \right\rangle. \qquad \square
\end{aligned}
$$

Proof of Lemma C.3.4.

$$
\begin{aligned}
&D_{\psi^{\mathcal{C}}}(\boldsymbol{x}^{(2)}, \boldsymbol{x}^{(0)}) - (1 + \eta\tau_0) D_{\psi^{\mathcal{C}}}(\boldsymbol{x}^{(2)}, \boldsymbol{x}^{(1)}) - D_{\psi^{\mathcal{C}}}(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(0)})\\
=&\big(\psi^{\mathcal{C}}(\boldsymbol{x}^{(2)}) - \psi^{\mathcal{C}}(\boldsymbol{x}^{(0)}) - \big\langle \nabla\psi^{\mathcal{C}}(\boldsymbol{x}^{(0)}), \boldsymbol{x}^{(2)} - \boldsymbol{x}^{(0)} \big\rangle\big)\\
&- (1 + \eta\tau_0) \big(\psi^{\mathcal{C}}(\boldsymbol{x}^{(2)}) - \psi^{\mathcal{C}}(\boldsymbol{x}^{(1)}) - \big\langle \nabla\psi^{\mathcal{C}}(\boldsymbol{x}^{(1)}), \boldsymbol{x}^{(2)} - \boldsymbol{x}^{(1)} \big\rangle\big)\\
&- \big(\psi^{\mathcal{C}}(\boldsymbol{x}^{(1)}) - \psi^{\mathcal{C}}(\boldsymbol{x}^{(0)}) - \big\langle \nabla\psi^{\mathcal{C}}(\boldsymbol{x}^{(0)}), \boldsymbol{x}^{(1)} - \boldsymbol{x}^{(0)} \big\rangle\big)\\
=&\eta\tau_0 \psi^{\mathcal{C}}(\boldsymbol{x}^{(1)}) - \eta\tau_0 \psi^{\mathcal{C}}(\boldsymbol{x}^{(2)}) + \big\langle (1 + \eta\tau_0) \nabla\psi^{\mathcal{C}}(\boldsymbol{x}^{(1)}) - \nabla\psi^{\mathcal{C}}(\boldsymbol{x}^{(0)}), \boldsymbol{x}^{(2)} - \boldsymbol{x}^{(1)} \big\rangle.
\end{aligned}
$$

Since

$$
\boldsymbol{x}^{(1)} = \operatorname*{argmin}_{\boldsymbol{x} \in \mathcal{C}} \left\{ \langle \boldsymbol{g}, \boldsymbol{x} \rangle + \tau_0 \psi^{\mathcal{C}}(\boldsymbol{x}) + \frac{1}{\eta} \left(\psi^{\mathcal{C}}(\boldsymbol{x}) - \psi^{\mathcal{C}}(\boldsymbol{x}^{(0)}) - \big\langle \nabla\psi^{\mathcal{C}}(\boldsymbol{x}^{(0)}), \boldsymbol{x} - \boldsymbol{x}^{(0)} \big\rangle \right) \right\},
$$

by first-order optimality, we have,

$$
\big\langle \eta \boldsymbol{g} + (1 + \eta\tau_0) \nabla\psi^{\mathcal{C}}(\boldsymbol{x}^{(1)}) - \nabla\psi^{\mathcal{C}}(\boldsymbol{x}^{(0)}), \boldsymbol{x}^{(2)} - \boldsymbol{x}^{(1)} \big\rangle \geq 0.
$$

Therefore,

$$
\begin{aligned}
&D_{\psi^{\mathcal{C}}}(\boldsymbol{x}^{(2)}, \boldsymbol{x}^{(0)}) - (1 + \eta\tau_0) D_{\psi^{\mathcal{C}}}(\boldsymbol{x}^{(2)}, \boldsymbol{x}^{(1)}) - D_{\psi^{\mathcal{C}}}(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(0)})\\
\geq&\eta\tau_0 \psi^{\mathcal{C}}(\boldsymbol{x}^{(1)}) - \eta\tau_0 \psi^{\mathcal{C}}(\boldsymbol{x}^{(2)}) + \eta \left\langle \boldsymbol{g}, \boldsymbol{x}^{(1)} - \boldsymbol{x}^{(2)} \right\rangle. \qquad \square
\end{aligned}
$$

C.4 Proof of Theorem 5.0.2

Theorem C.4.1 (Formal Version of Theorem 5.0.2)**.** Consider Algorithm 1. When $\frac{\eta^{\mathrm{anc}}_s}{\eta_s} \leq \tau C^{\eta,T}_s$ for any $s \in \mathcal{S}$, where $C^{\eta,T}_s := \frac{\gamma^2 \sum_{h \in s} \mu_c(h)}{2C^-_s \left(\log T + \log|\mathcal{S}| + \log\frac{1}{\delta}\right)}$, and (A), (B), (C) are satisfied, we

have the following guarantee with probability $1-2\delta$.

$$\begin{aligned}
&\sum_{t=2}^{T} D_{\psi^{\Pi}}(\boldsymbol{\mu}^{(\tau,\gamma),*}, \boldsymbol{\mu}^{\overline{\boldsymbol{\pi}}^{(t)}}) \\
\leq &\frac{4C_{\text{visit}}}{\tau} \sum_{s\in\mathcal{S}} C_s^{\text{diff}} \mu^{(\tau,\gamma),*}(\sigma(s)) \left\|\boldsymbol{q}\right\|_{\infty} \eta_s M_2 T \\
&+ \frac{2C_{\text{visit}}}{\tau} \sum_{s\in\mathcal{S}} \frac{1}{\eta_s \mu_{p(s)}^{(t_s^1)}(\sigma(s))} \mu^{(\tau,\gamma),*}(\sigma(s)) \max_{\boldsymbol{x},\boldsymbol{y}\in\Delta_{\gamma_s,\boldsymbol{\nu}_s}^{|\mathcal{A}_s|}} D_{\psi_s^{\Delta}}(\boldsymbol{x},\boldsymbol{y}) \\
&+ \frac{4C_{\text{visit}}^2}{\tau} \sum_{s\in\mathcal{S}} \frac{1}{\eta_s} \mu^{(\tau,\gamma),*}(\sigma(s)) \max_{\boldsymbol{x},\boldsymbol{y}\in\Delta_{\gamma_s,\boldsymbol{\nu}_s}^{|\mathcal{A}_s|}} D_{\psi_s^{\Delta}}(\boldsymbol{x},\boldsymbol{y}) \qquad \text{(C.4.1)} \\
&+ \frac{4C_{\text{visit}}}{\tau} \left(1+\frac{1}{\gamma}\right) (\left\|\boldsymbol{q}\right\|_{\infty} + \tau\psi^{\max}) \sqrt{2T \log \frac{|\mathcal{S}|}{\delta}} + \sum_{s\in\mathcal{S}} \frac{4C_{\text{visit}}^2}{\eta_s \tau} \max_{\boldsymbol{x},\boldsymbol{y}\in\Delta_{\gamma_s,\boldsymbol{\nu}_s}^{|\mathcal{A}_s|}} D_{\psi_s^{\Delta}}(\boldsymbol{x},\boldsymbol{y}),
\end{aligned}$$

where $C_{\text{visit}} \coloneqq \frac{\log T + \log|\mathcal{S}| + \log\frac{1}{\delta}}{\gamma^2 \min_{s\in\mathcal{S}} \sum_{h\in s} \mu_c(h)}$ is the maximum gap between any two adjacent visits to an infoset.

Proof Sketch. (i). Firstly, we show that the estimates in Algorithm 1 are unbiased so that the conditions of Lemma 5.1.3 are met. (ii). By Lemma 5.1.3 and union bound, we can extend the result of full information feedback to the stochastic feedback. (iii). To ensure the coefficient for the cumulated distance to NE after telescoping is still positive, we need to bound the largest gap between the timesteps of two consecutive visits of any infoset.

For any infoset $s \in \mathcal{S}$, we define $T_s \coloneqq \{t_s^1, t_s^2, \cdots\}$, where each $t_s^k \in [T]$ is the timestep that s is along the sampled trajectory. Then, we will show Algorithm 1 uses unbiased estimators so that we can derive an upper-bound by Lemma 5.1.3. Note that for any $\boldsymbol{u} \in \Delta_{\gamma_s,\boldsymbol{\nu}_s}^{|\mathcal{A}_s|}$, the expectation of the additional regularizer term is,

$$\begin{aligned}
\Pr\left(t \in T_s\right) \left(\frac{\tau}{\mu_{p(s)}^{(t)}(\sigma(s))} \psi_s^{\Delta}(\boldsymbol{u})\right) =& \mu_{p(s)}^{(t)}(\sigma(s)) \mu_{-p(s)}^{(t)}(s) \left(\frac{\tau}{\mu_{p(s)}^{(t)}(\sigma(s))} \psi_s^{\Delta}(\boldsymbol{u})\right) \\
=& \tau \mu_{-p(s)}^{(t)}(s) \psi_s^{\Delta}(\boldsymbol{u}).
\end{aligned}$$

Let $s(h)$ denote the infoset that the node h is in. For the original utility, suppose $p(s) = 1$ without loss of generality, the expectation of $\widetilde{q}^{(t)}(s,a)$ for any $a \in \mathcal{A}_s$ is,

$$\begin{aligned}
&\frac{1}{\pi_1^{(t)}(a \mid s)} \sum_{h'\in\mathcal{H}:\ \exists h\in s,(h,a)\sqsubseteq h'} \mu_1^{(t)}(\sigma_1(h')) \mu_2^{(t)}(\sigma_2(h')) \mu_c(h') \mathcal{U}_1(h') \\
&- \frac{1}{\pi_1^{(t)}(a \mid s)} \sum_{h'\in\mathcal{H}_1:\ \exists h\in s,(h,a)\sqsubseteq h'} \mu_1^{(t)}(\sigma_1(h')) \mu_2^{(t)}(\sigma_2(h')) \mu_c(h') \psi_{s(h)}^{\Delta}(\pi_1^{(t)}) \\
&+ \frac{1}{\pi_1^{(t)}(a \mid s)} \sum_{h'\in\mathcal{H}_2:\ \exists h\in s,(h,a)\sqsubseteq h'} \mu_1^{(t)}(\sigma_1(h')) \mu_2^{(t)}(\sigma_2(h')) \mu_c(h') \psi_{s(h)}^{\Delta}(\pi_2^{(t)}),
\end{aligned}$$

which is equal to $q^{(t)}(s,a)$ by definition.

By Lemma 5.1.3 and union bound, with probability at least $1-\delta$, the following is satisfied for all infosets $s \in \mathcal{S}$,

$$\Pr\left(t \in T_s\right)\left(\frac{\tau}{\mu^{(t)}_{p(s)}(\sigma(s))}\psi^{\Delta}_s(\boldsymbol{u})\right) = \mu^{(t)}_{p(s)}(\sigma(s))\mu^{(t)}_{-p(s)}(s)\left(\frac{\tau}{\mu^{(t)}_{p(s)}(\sigma(s))}\psi^{\Delta}_s(\boldsymbol{u})\right) = \tau\mu^{(t)}_{-p(s)}(s)\psi^{\Delta}_s(\boldsymbol{u}).$$

where $\psi^{\max}$ is the upperbound of ψ^{Δ}_s, which is $\max_{s\in\mathcal{S}}\frac{\alpha_s}{2}$ when it is Euclidean norm and $\max_{s\in\mathcal{S}}\alpha_s\log|\mathcal{A}_s|$ when it is entropy.

Let's define $t^0_s = 1$ for notational simplicity. Similar to the proof of Theorem 5.1.2, for each $k \le |T_s|-1$, we have

$$\begin{aligned}
&\eta_s\tau\psi^{\Delta}_s(\pi^{(t^k_s)}_{p(s)}(\cdot\,|\,s)) - \eta_s\tau\psi^{\Delta}_s(\pi_{p(s)}(\cdot\,|\,s)) + \eta_s\left\langle -\widetilde{q}^{(t^k_s)}(s,\cdot), \pi^{(t^k_s)}_{p(s)}(\cdot\,|\,s) - \pi_{p(s)}(\cdot\,|\,s)\right\rangle\\
\le& D_{\psi^{\Delta}_s}(\pi_{p(s)}(\cdot\,|\,s), \overline{\pi}^{(t^k_s)}_{p(s)}(\cdot\,|\,s)) - (1+\eta_s\tau)D_{\psi^{\Delta}_s}(\pi_{p(s)}(\cdot\,|\,s), \overline{\pi}^{(t^{k+1}_s)}_{p(s)}(\cdot\,|\,s))\\
&- (1+\eta_s\tau)D_{\psi^{\Delta}_s}(\overline{\pi}^{(t^{k+1}_s)}_{p(s)}(\cdot\,|\,s), \pi^{(t^k_s)}_{p(s)}(\cdot\,|\,s))\\
&+ \eta_s\left\langle \widetilde{q}^{(t^{k-1}_s)}(s,\cdot) - \widetilde{q}^{(t^k_s)}(s,\cdot), \pi^{(t^k_s)}_{p(s)}(\cdot\,|\,s) - \overline{\pi}^{(t^{k+1}_s)}_{p(s)}(\cdot\,|\,s)\right\rangle.
\end{aligned}$$

By multiplying $m^{(t^k_s)}_s = \frac{1}{\mu^{(t^k_s)}_{p(s)}(\sigma(s))}$ on both sides and telescoping, we have

$$\begin{aligned}
&\sum_{k=1}^{|T_s|-1}\left(\frac{\tau}{\mu^{(t^k_s)}_{p(s)}(\sigma(s))}\left(\psi^{\Delta}_s(\pi^{(t^k_s)}_{p(s)}(\cdot\,|\,s)) - \psi^{\Delta}_s(\pi_{p(s)}(\cdot\,|\,s))\right)\right.\\
&\left.+\frac{1}{\mu^{(t^k_s)}_{p(s)}(\sigma(s))}\langle -\widetilde{q}^{(t^k_s)}(s,\cdot), \pi^{(t^k_s)}_{p(s)}(\cdot\,|\,s) - \pi_{p(s)}(\cdot\,|\,s)\rangle\right)\\
\le&\sum_{k=2}^{|T_s|}\left(\frac{1}{\mu^{(t^k_s)}_{p(s)}(\sigma(s))} - \frac{1}{\mu^{(t^{k-1}_s)}_{p(s)}(\sigma(s))} - \frac{\eta_s\tau}{\mu^{(t^{k-1}_s)}_{p(s)}(\sigma(s))}\right)D_{\psi^{\Delta}_s}(\pi_{p(s)}(\cdot\,|\,s), \overline{\pi}^{(t^k_s)}_{p(s)}(\cdot\,|\,s))\\
&+ 2C^{\mathrm{diff}}_s\left\|\boldsymbol{q}\right\|_{\infty}\eta_s M_2|T_s| + \frac{1}{\mu^{(t^1_s)}_{p(s)}(\sigma(s))}D_{\psi^{\Delta}_s}(\pi_{p(s)}(\cdot\,|\,s), \overline{\pi}^{(t^1_s)}_{p(s)}(\cdot\,|\,s)).
\end{aligned}$$

Since the probability of visiting infoset s at timestep t is at least $\gamma^2\sum_{h\in s}\mu_c(h)$,

$$\Pr\left(|t^k_s - t^{k-1}_s| > K_s\right) \le (1-\gamma^2\sum_{h\in s}\mu_c(h))^{K_s} \le \exp(-\gamma^2\sum_{h\in s}\mu_c(h)K_s).$$

Therefore, with probability $1-\delta$, all infosets $s\in\mathcal{S}$ satisfies that for any $2 \le k \le |T_s|$, $|t^k_s - t^{k-1}_s| \le \frac{\log T + \log|\mathcal{S}| + \log\frac{1}{\delta}}{\gamma^2\sum_{h\in s}\mu_c(h)} =: K_s$. Then,

$$\begin{aligned}
\frac{1}{\mu^{(t^k_s)}_{p(s)}(\sigma(s))} - \frac{1}{\mu^{(t^{k-1}_s)}_{p(s)}(\sigma(s))} - \frac{\eta_s\tau}{\mu^{(t^{k-1}_s)}_{p(s)}(\sigma(s))} \le& C^{-}_s\eta^{\mathrm{anc}}_s\frac{\log T + \log|\mathcal{S}| + \log\frac{1}{\delta}}{\gamma^2\sum_{h\in s}\mu_c(h)} - \eta_s\tau\\
=& C^{-}_s\eta^{\mathrm{anc}}_s\frac{\log T + \log|\mathcal{S}| + \log\frac{1}{\delta}}{\gamma^2\sum_{h\in s}\mu_c(h)} - \eta_s\tau.
\end{aligned}$$

Therefore, when $\frac{\eta_s^{\mathrm{anc}}}{\eta_s} \le \frac{\tau\gamma^2 \sum_{h\in s}\mu_c(h)}{2C_s^-\left(\log T+\log|\mathcal{S}|+\log\frac{1}{\delta}\right)}$, the inequality above is upper-bounded by $-\frac{\eta_s\tau}{2}$. Moreover, we can write it as (let $t_s^{|T_s|+1} = T+1$ for notational simplicity),

$$
\begin{aligned}
&\sum_{k=2}^{|T_s|}\left(\frac{1}{\mu_{p(s)}^{(t_s^k)}(\sigma(s))}-\frac{1}{\mu_{p(s)}^{(t_s^{k-1})}(\sigma(s))}-\frac{\eta_s\tau}{\mu_{p(s)}^{(t_s^{k-1})}(\sigma(s))}\right) D_{\psi_s^\Delta}(\pi_{p(s)}(\cdot\,|\,s),\bar{\pi}_{p(s)}^{(t_s^k)}(\cdot\,|\,s))\\
\le& -\sum_{k=2}^{|T_s|}\sum_{t=t_s^{(k)}}^{t_s^{(k+1)}-1}\frac{\eta_s\tau}{2K_s} D_{\psi_s^\Delta}(\pi_{p(s)}(\cdot\,|\,s),\bar{\pi}_{p(s)}^{(t_s^k)}(\cdot\,|\,s))\\
=& -\frac{\eta_s\tau\gamma^2\sum_{h\in s}\mu_c(h)}{2\left(\log T+\log|\mathcal{S}|+\log\frac{1}{\delta}\right)}\sum_{k=2}^{|T_s|}\sum_{t=t_s^{(k)}}^{t_s^{(k+1)}-1} D_{\psi_s^\Delta}(\pi_{p(s)}(\cdot\,|\,s),\bar{\pi}_{p(s)}^{(t_s^k)}(\cdot\,|\,s))\\
\overset{(i)}{=}& -\frac{\eta_s\tau\gamma^2\sum_{h\in s}\mu_c(h)}{2\left(\log T+\log|\mathcal{S}|+\log\frac{1}{\delta}\right)}\sum_{t=t_s^{(2)}}^{T} D_{\psi_s^\Delta}(\pi_{p(s)}(\cdot\,|\,s),\bar{\pi}_{p(s)}^{(t)}(\cdot\,|\,s))\\
\le& -\frac{\eta_s\tau\gamma^2\sum_{h\in s}\mu_c(h)}{2\left(\log T+\log|\mathcal{S}|+\log\frac{1}{\delta}\right)}\sum_{t=2}^{T} D_{\psi_s^\Delta}(\pi_{p(s)}(\cdot\,|\,s),\bar{\pi}_{p(s)}^{(t)}(\cdot\,|\,s)) + 2K_s\max_{\boldsymbol{x},\boldsymbol{y}\in\Delta_{\gamma_s,\boldsymbol{\nu}_s}^{|\mathcal{A}_s|}} D_{\psi_s^\Delta}(\boldsymbol{x},\boldsymbol{y}).
\end{aligned}
$$

(i) uses the fact that for any $t\in[t_s^k, t_s^{k+1}-1]$, $\bar{\pi}_{p(s)}^{(t)}(\cdot\,|\,s) = \bar{\pi}_{p(s)}^{(t_s^k)}(\cdot\,|\,s)$.

Then, following rest of the proof for Theorem 5.1.2, we finish the proof. □

C.4.1 Proof of Lemma 5.1.3

Let $d^{(t)}(\boldsymbol{u}) \coloneqq \left(f^{(t)}(\boldsymbol{u}) - f^{(t)}(\boldsymbol{u}^{(t)})\right) - \left(\widetilde{f}^{(t)}(\boldsymbol{u}) - \widetilde{f}^{(t)}(\boldsymbol{u}^{(t)})\right)$. By the property of $f^{(t)}$, since $\boldsymbol{u}^{(t)}$ is deterministically influenced by $\widetilde{f}^{(1)}, \widetilde{f}^{(2)}, \cdots, \widetilde{f}^{(t-1)}$, $\mathbb{E}\left[\widetilde{f}^{(t)}(\boldsymbol{u}^{(t)})\,|\,\widetilde{f}^{(1)}, \widetilde{f}^{(2)}, \cdots, \widetilde{f}^{(t-1)}\right] = f^{(t)}(\boldsymbol{u}^{(t)})$. Moreover, $\mathbb{E}\left[\widetilde{f}^{(t)}(\boldsymbol{u})\,|\,\widetilde{f}^{(1)}, \widetilde{f}^{(2)}, \cdots, \widetilde{f}^{(t-1)}\right] = f^{(t)}(\boldsymbol{u})$ for any fixed $\boldsymbol{u}\in\mathcal{C}$. Therefore, $d^{(t)}(\boldsymbol{u})$ is a martingale difference sequence. Then, we can apply the Azuma-Hoeffding inequality in the following.

Lemma C.4.2 (Azuma-Hoeffding inequality)**.** For any martingale difference sequence $x^{(1)}, x^{(2)}, \cdots, x^{(T)}$ with $x^{(t)}\in[a^{(t)}, b^{(t)}]$, we have

$$
\Pr\left(\sum_{t=1}^{T} x^{(t)} \ge w\right) \le \exp\left(-\frac{2w^2}{\sum_{t=1}^{T}\left(b^{(t)}-a^{(t)}\right)^2}\right).
$$

Note that $\left|d^{(t)}(\boldsymbol{u})\right| \le M + \widetilde{M}$. By applying Lemma C.4.2, we have

$$
\Pr\left(\sum_{t=1}^{T} d^{(t)}(\boldsymbol{u}) \ge w\right) \le \exp\left(-\frac{2w^2}{\sum_{t=1}^{T} 4(M+\widetilde{M})^2}\right).
$$

Therefore, when taking $w = (M + \widetilde{M})\sqrt{2T \log \frac{1}{\delta}}$, with probability at least $1 - \delta$,

$$\sum_{t=1}^{T} \left(f^{(t)}(\boldsymbol{u}) - f^{(t)}(\boldsymbol{u}^{(t)})\right) \leq \sum_{t=1}^{T} \left(\widetilde{f}^{(t)}(\boldsymbol{u}) - \widetilde{f}^{(t)}(\boldsymbol{u}^{(t)})\right) + (M + \widetilde{M})\sqrt{2T \log \frac{1}{\delta}}. \qquad \square$$

C.5 Auxiliary Lemmas

In this section, we present the auxiliary lemmas for the theorems proved in the previous part.

C.5.1 Upperbound of Feedback

Lemma C.5.1 (Upperbound of Feedback $q^{(t)}(s, \cdot)$)**.** Consider the update-rule (5.1.2). For any timestep $t \in [T]$, we have the following upper-bound on $q^{(t)}(s, \cdot)$ and its unbiased estimator $\widetilde{q}^{(t)}(s, \cdot)$, no matter whether it is counterfactual value, trajectory Q-value, or Q-value.

$$\|\boldsymbol{q}\|_\infty \coloneqq \begin{cases} \frac{\frac{\tau}{M_1}\|\boldsymbol{\alpha}\|_\infty \mathcal{D}\psi^{\max}+1}{\min_{s\in\mathcal{S}, a\in\mathcal{A}_s} \gamma_s \nu_{s,a}} & \text{Outcome Sampling of Trajectory Q-value} \\ \frac{\tau}{M_1}\|\boldsymbol{\alpha}\|_\infty \mathcal{D}\psi^{\max} + 1 & \text{Otherwise} \end{cases} \tag{C.5.1}$$

where $\mathcal{D} \coloneqq \max_{h\in\mathcal{H}} \mathcal{D}(h)$ is the maximum depth of infoset and $\psi^{\max}$ is the maximum of the regularizer, which is $\frac{1}{2\min_{s\in\mathcal{S}}|\mathcal{A}_s|}$ for Euclidean distance and $\max_{s\in\mathcal{S}} \log|\mathcal{A}_s|$ for entropy.

Proof. When calculating the feedback $\widetilde{q}^{(t)}(s, a)$ for outcome sampling of trajectory Q-value, we need to divide the probability of choosing action a, which is $\pi^{(t)}(a \,|\, s) \geq \min_{s\in\mathcal{S}, a\in\mathcal{A}_s} \gamma_s \nu_{s,a}$. Then, its upperbound is that of the full-information feedback setting divided by the constant $\min_{s\in\mathcal{S}, a\in\mathcal{A}_s} \gamma_s \nu_{s,a}$. Therefore, in the following, we will focus on the upperbound of $q^{(t)}(s, \cdot)$ in the full-information feedback setting.

In the following proof, we only consider $s \in \mathcal{S}_1$ since player 1, 2 are symmetric. Furthermore, we only need to prove the upper-bound above when $q^{(t)}(s, \cdot)$ is Q-value, since by definition, the Q-value $Q_i^{\boldsymbol{\pi}}(s, a) = \frac{\overline{Q}_i^{\boldsymbol{\pi}}(s,a)}{\sum_{h\in s} \mu_c(h)\mu_1^{\pi_1}(\sigma_1(h))\mu_2^{\pi_2}(\sigma_2(h))} \geq \overline{Q}_i^{\boldsymbol{\pi}}(s, a)$ (similarly, it is also larger than the counterfactual value).

Let $s(h)$ be the infoset that node h is in. Firstly, when $\tau = 0$, which means only considering the contribution of $\mathcal{U}_1$ to $q^{(t)}(s, \cdot)$, for every $s \in \mathcal{S}_1$, we have

$$\begin{aligned}
&|Q_i^{\boldsymbol{\pi}}(s, a)| \\
=&\frac{1}{\sum_{h\in s} \mu_c(h)\mu_1^{\pi_1}(\sigma_1(h))\mu_2^{\pi_2}(\sigma_2(h))} \left| \sum_{h' \colon \exists h\in s, (h,a)\sqsubseteq h'} \mu_c(h')\mathcal{U}_1(h')\mu_1^{\pi_1}(\sigma_1(h'))\mu_2^{\pi_2}(\sigma_2(h')) \right| \\
\overset{(i)}{=}&\frac{1}{\sum_{h\in s} \mu_c(h)\mu_1^{\pi_1}(\sigma_1(h))\mu_2^{\pi_2}(\sigma_2(h))} \left| \sum_{h' \colon \exists h\in s, (h,a)\sqsubseteq h', \mathcal{A}_{s(h')}=\emptyset} \mu_c(h')\mathcal{U}_1(h')\mu_1^{\pi_1}(\sigma_1(h'))\mu_2^{\pi_2}(\sigma_2(h')) \right| \\
\leq&\frac{1}{\sum_{h\in s} \mu_c(h)\mu_1^{\pi_1}(\sigma_1(h))\mu_2^{\pi_2}(\sigma_2(h))} \sum_{h' \colon \exists h\in s, (h,a)\sqsubseteq h', \mathcal{A}_{s(h')}=\emptyset} \mu_c(h')\mu_1^{\pi_1}(\sigma_1(h'))\mu_2^{\pi_2}(\sigma_2(h')) \\
\overset{(ii)}{=}&\frac{1}{\sum_{h\in s} \mu_c(h)\mu_1^{\pi_1}(\sigma_1(h))\mu_2^{\pi_2}(\sigma_2(h))} \sum_{h\in s} \mu_c(h)\mu_1^{\pi_1}(\sigma_1(h))\mu_2^{\pi_2}(\sigma_2(h)) = 1.
\end{aligned}$$

(i) is because $\mathcal{U}_1(h) \neq 0$ only if h is a terminal node. (ii) is by the tree structure of EFG. Now consider $\tau > 0$ and $\mathcal{U}_1(h) \equiv 0$ for any $h \in \mathcal{H}$. Moreover, we will only show the upperbound when using dilated regularizer, since bidilated regularizer is upperbounded by the dilated one.

Let $S^{(t)}(s) = \left\langle q^{(t)}(s,\cdot), \pi^{(t)}_{p(s)}(\cdot \,|\, s) \right\rangle - \frac{\tau}{m_s^{(t)}} \psi_s^{\Delta}(\pi^{(t)}_{p(s)}(\cdot \,|\, s))$ when $\mathcal{A}_s \neq \emptyset$ (s is not terminal node) and $S^{(t)}(s) = 0$ when $\mathcal{A}_s = \emptyset$ (s is the terminal node).

We will prove $\left|S^{(t)}(s)\right| \leq \frac{\tau}{M_1} \left\|\boldsymbol{\alpha}\right\|_\infty (\mathcal{D} - \mathcal{D}(s)) \, \psi^{\max}$ by induction. For infoset $s \in \mathcal{S}_1$ with $\mathcal{D}(s) = D$, we have $S^{(t)}(s) = 0 = \tau \left\|\boldsymbol{\alpha}\right\|_\infty (\mathcal{D} - \mathcal{D}(s)) \, \psi^{\max}$. Therefore, the initial step of induction is completed.

Consider when all $s' \in \mathcal{S}_1$ with $\mathcal{D}(s') > d$ for some constant d, $S^{(t)}(s') \leq \frac{\tau}{M_1} \left\|\boldsymbol{\alpha}\right\|_\infty (\mathcal{D} - \mathcal{D}(s')) \, \psi^{\max}$. Let $\Pr(h \to h')$ be the probability of reaching h' from node h, when considering all nodes encountered along the path (player $1, 2$ action node and the chance node) for notational simplicity. Then, for infoset $s \in \mathcal{S}_1$ with $\mathcal{D}(s) = d$, we have

$$
\begin{aligned}
&\left|Q_i^{\boldsymbol{\pi}}(s,a)\right| \\
=&\left| \sum_{s' \in \mathcal{S}_1 \,:\, \sigma(s') = (s,a)} S^{(t)}(s') \sum_{h \in s} \frac{\mu_c(h) \mu_1^{\pi_1}(\sigma_1(h)) \mu_2^{\pi_2}(\sigma_2(h))}{\sum_{h' \in s} \mu_c(h') \mu_1^{\pi_1}(\sigma_1(h')) \mu_2^{\pi_2}(\sigma_2(h'))} \sum_{h' \in s'} \Pr(h \to h') \right| \\
\leq& \sum_{s' \in \mathcal{S}_1 \,:\, \sigma(s') = (s,a)} \frac{\tau}{M_1} \left\|\boldsymbol{\alpha}\right\|_\infty (\mathcal{D} - \mathcal{D}(s) - 1) \, \psi^{\max} \\
&\cdot \sum_{h \in s} \frac{\mu_c(h) \mu_1^{\pi_1}(\sigma_1(h)) \mu_2^{\pi_2}(\sigma_2(h))}{\sum_{h' \in s} \mu_c(h') \mu_1^{\pi_1}(\sigma_1(h')) \mu_2^{\pi_2}(\sigma_2(h'))} \sum_{h' \in s'} \Pr(h \to h') \\
=&\frac{\tau}{M_1} \left\|\boldsymbol{\alpha}\right\|_\infty (\mathcal{D} - \mathcal{D}(s) - 1) \, \psi^{\max} \\
&\cdot \sum_{h \in s} \frac{\mu_c(h) \mu_1^{\pi_1}(\sigma_1(h)) \mu_2^{\pi_2}(\sigma_2(h))}{\sum_{h' \in s} \mu_c(h') \mu_1^{\pi_1}(\sigma_1(h')) \mu_2^{\pi_2}(\sigma_2(h'))} \sum_{h' \in \mathcal{H}_1 \,:\, \sigma_1(h') = (s,a)} \Pr(h \to h') \\
=&\frac{\tau}{M_1} \left\|\boldsymbol{\alpha}\right\|_\infty (\mathcal{D} - \mathcal{D}(s) - 1) \, \psi^{\max}.
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
\left|S^{(t)}(s)\right| &= \left| \left\langle q^{(t)}(s,\cdot), \pi^{(t)}_{p(s)}(\cdot \,|\, s) \right\rangle - \frac{\tau}{m_s^{(t)}} \psi_s^{\Delta}(\pi^{(t)}_{p(s)}(\cdot \,|\, s)) \right| \\
\leq& \frac{\tau}{M_1} \left\|\boldsymbol{\alpha}\right\|_\infty (\mathcal{D} - \mathcal{D}(s) - 1) \, \psi^{\max} + \frac{\tau}{M_1} \left\|\boldsymbol{\alpha}\right\|_\infty \psi^{\max} \\
=& \frac{\tau}{M_1} \left\|\boldsymbol{\alpha}\right\|_\infty (\mathcal{D} - \mathcal{D}(s)) \, \psi^{\max}.
\end{aligned}
$$

This concludes the induction step. □

C.5.2 Bounding M_1, M_2

By choosing γ_s as a fixed constant $\gamma_0 > 0$ for any player $i \in [2]$ and $s \in \mathcal{S}_i$. $\nu_{s,a}$ is chosen to be proportional to the number of terminal infosets ($s' \in \mathcal{S}$ with $\mathcal{A}_{s'} = \emptyset$) in the subtree rooted at (s,a), we can get $\gamma \geq \frac{\gamma_0^{\mathcal{D}}}{|\mathcal{S}|}$. Then, we have the following lowerbound and upperbound on $m_s^{(t)}$.

Lemma C.5.2. When $\mu_1^{(t)}, \mu_2^{(t)} \succeq \gamma$, $m_s^{(t)}$ are lowerbounded by M_1 and upperbounded by M_2 with the following M_1, M_2 for different feedback.

$$M_1 \coloneqq \begin{cases} \gamma \min_{s\in\mathcal{S}} \sum_{h\in s} \mu_c(h) & \text{Q-value} \\ 1 & \text{Trajectory Q-value} \\ 1 & \text{Counterfactual Value} \end{cases} \quad M_2 \coloneqq \begin{cases} 1 & \text{Q-value} \\ \frac{1}{\gamma} & \text{Trajectory Q-value} \\ 1 & \text{Counterfactual Value} \end{cases} \tag{C.5.2}$$

Proof. We only prove the lowerbound and upperbound for infoset $s \in \mathcal{S}_1$ since two players are symmetric.

For counterfactual value, since $m_s^{(t)} \equiv 1$, $M_1, M_2 = 1$.

For trajectory Q-value, since $m_s^{(t)} = \frac{1}{\mu_1^{(t)}(\sigma(s))}$, we have $M_2 = \frac{1}{\gamma} \geq m_s^{(t)} \geq 1 = M_1$.

For Q-value, $m_s^{(t)} = \sum_{h\in s} \mu_c(h)\mu_2^{(t)}(\sigma_2(h))$. Since the reach probability $\mu_1^{\pi_1}(\sigma(s)) \sum_{h\in s} \mu_c(h)\mu_2^{(t)}(\sigma_2(h)) \leq 1$ for any π_1, we can let π_1 play deterministically to reach s. In this way, $m_s^{(t)}$ is equal to the reach probability so it is also upperbounded by one. At the same time, $m_s^{(t)} \geq \gamma \sum_{h\in s} \mu_c(h) \geq \gamma \min_{s\in\mathcal{S}} \sum_{h\in s} \mu_c(h)$. □

C.5.3 Proof of Lemma C.2.2

Lemma C.5.3. Consider update-rule (5.1.2). When $\psi_s^\Delta(\boldsymbol{u}) = \frac{\alpha_s}{2} \sum_{a\in\mathcal{A}_s} u_a^2$ is the Euclidean distance where $\alpha_s > 0$ is a constant, we have

$$C_s^{\text{diff}} = \frac{|\mathcal{A}_s|}{\alpha_s} \|\boldsymbol{q}\|_\infty + \frac{\tau}{M_1}\sqrt{|\mathcal{A}_s|}. \tag{C.5.3}$$

Let $\tau_s^{(t)} \coloneqq \frac{\tau \mu_{-p(s)}^{(t)}(s)}{m_s^{(t)}}$. When $\psi_s^\Delta(\boldsymbol{u}) = \frac{\alpha_s}{2} \sum_{a\in\mathcal{A}_s} u_a^2$, we have

$$\begin{aligned}
&\left\|\pi_{p(s)}^{(t)}(\cdot \mid s) - \bar{\pi}_{p(s)}^{(t)}(\cdot \mid s)\right\| \\
=&\left\|\text{Proj}_{\Delta_{\gamma_s,\nu_s}^{|\mathcal{A}_s|}}\left(\frac{\bar{\pi}_{p(s)}^{(t)}(\cdot \mid s)}{1+\eta_s\tau_s^{(t-1)}} + \frac{\eta_s}{\alpha_s(1+\eta_s\tau_s^{(t-1)})} q^{(t-1)}(s,\cdot)\right) - \bar{\pi}_{p(s)}^{(t)}(\cdot \mid s)\right\| \\
\leq&\left\|\frac{\bar{\pi}_{p(s)}^{(t)}(\cdot \mid s)}{1+\eta_s\tau_s^{(t-1)}} + \frac{\eta_s}{\alpha_s(1+\eta_s\tau_s^{(t-1)})} q^{(t-1)}(s,\cdot) - \bar{\pi}_{p(s)}^{(t)}(\cdot \mid s)\right\| \\
\leq&\eta_s\left(\frac{1}{\alpha_s(1+\eta_s\tau_s^{(t-1)})}\left\|q^{(t-1)}(s,\cdot)\right\| + \frac{\tau_s^{(t-1)}}{1+\eta_s\tau_s^{(t-1)}}\left\|\bar{\pi}_{p(s)}^{(t)}(\cdot \mid s)\right\|\right) \\
\leq&\eta_s\left(\frac{1}{\alpha_s}\left\|q^{(t-1)}(s,\cdot)\right\| + \tau_s^{(t-1)}\left\|\bar{\pi}_{p(s)}^{(t)}(\cdot \mid s)\right\|\right) \\
\leq&\eta_s\left(\frac{\sqrt{|\mathcal{A}_s|}}{\alpha_s}\|\boldsymbol{q}\|_\infty + \frac{\tau}{M_1}\right).
\end{aligned}$$

In the last line, we use the fact that $\mu_{-p(s)}^{(t)}(s) \leq 1$.

As a result, $\left\|\pi^{(t)}_{p(s)}(\cdot\,|\,s)-\bar{\pi}^{(t)}_{p(s)}(\cdot\,|\,s)\right\|_1 \le \eta_s\left(\frac{|\mathcal{A}_s|}{\alpha_s}\|\boldsymbol{q}\|_\infty+\frac{\tau}{M_1}\sqrt{|\mathcal{A}_s|}\right)$.
Similarly, $\left\|\bar{\pi}^{(t+1)}_{p(s)}(\cdot\,|\,s)-\bar{\pi}^{(t)}_{p(s)}(\cdot\,|\,s)\right\|_1 \le \eta_s\left(\frac{|\mathcal{A}_s|}{\alpha_s}\|\boldsymbol{q}\|_\infty+\frac{\tau}{M_1}\sqrt{|\mathcal{A}_s|}\right)$. □

Proof of Lemma C.2.5. Let $\tau_s^{(t)} := \frac{\tau\mu^{(t)}_{-p(s)}(s)}{m_s^{(t)}}$. When $\psi_s^{\Delta}(\boldsymbol{u}) = \alpha_s\left(\log|\mathcal{A}_s| + \sum_{a\in\mathcal{A}_s} u_a\log u_a\right)$, the update-rule (5.1.2) is equivalent to

$$\begin{aligned}
&\pi^{(t)}(a\,|\,s)\\
=&\max\left\{\frac{\bar{\pi}^{(t)}(a\,|\,s)^{\frac{1}{1+\eta_s\tau_s^{(t-1)}}}\exp\left(\frac{\eta_s}{\alpha_s\left(1+\eta_s\tau_s^{(t-1)}\right)}q^{(t)}(s,a)\right)}{Z},\gamma_s\nu_{s,a}\right\}\\
=&\max\left\{\frac{\bar{\pi}^{(t)}(a\,|\,s)\exp\left(\frac{\eta_s}{\alpha_s\left(1+\eta_s\tau_s^{(t-1)}\right)}q^{(t)}(s,a)-\frac{\eta_s\tau}{m_s^{(t-1)}\left(1+\eta_s\tau_s^{(t-1)}\right)}\log\bar{\pi}^{(t)}(a\,|\,s)\right)}{Z},\gamma_s\nu_{s,a}\right\}
\end{aligned}$$

for any $a\in\mathcal{A}_s$, where $Z>0$ is a normalizing constant to ensure $\pi^{(t)}(a\,|\,s)$ is still a probability distribution over $\Delta^{|\mathcal{A}_s|}$. The equivalency is proved in Lemma C.5.4. For notational simplicity, we define $l^{(t)}(s,a) := -\frac{1}{\alpha_s\left(1+\eta_s\tau_s^{(t-1)}\right)}q^{(t)}(s,a)+\frac{\tau}{m_s^{(t-1)}\left(1+\eta_s\tau_s^{(t-1)}\right)}\log\bar{\pi}^{(t)}(a\,|\,s)$ so that $\pi^{(t)}(a\,|\,s)=\max\left\{\frac{\bar{\pi}^{(t)}(a\,|\,s)\exp\left(-\eta_s l^{(t)}(s,a)\right)}{Z},\gamma_s\nu_{s,a}\right\}$.

Firstly, for $\gamma_s=1$, we have $\pi^{(t)}_{p(s)}(\cdot\,|\,s)=\bar{\pi}^{(t)}_{p(s)}(\cdot\,|\,s)=\boldsymbol{\nu}_s$. Therefore, $C_s^{\text{diff}}=0$ and $\frac{\pi^{(t)}(a\,|\,s)}{\bar{\pi}^{(t)}(a\,|\,s)}=1$ for any $a\in\mathcal{A}_s$. In the following, we assume $\gamma_s<1$.

We can see that $\pi^{(t)}(a\,|\,s)$ is monotonically decreasing with respect to Z. When $Z<\exp\left(-\eta_s\max_{a'\in\mathcal{A}_s}l^{(t)}(s,a')\right)$, for any $a\in\mathcal{A}_s$, we have $\pi^{(t)}(a\,|\,s)\ge\frac{\bar{\pi}^{(t)}(a\,|\,s)\exp\left(-\eta_s l^{(t)}(s,a)\right)}{Z}>\bar{\pi}^{(t)}(a\,|\,s)$. Then, $\sum_{a\in\mathcal{A}_s}\pi^{(t)}(a\,|\,s)>1$.

Therefore, $Z\ge\exp\left(-\eta_s\max_{a'\in\mathcal{A}_s}l^{(t)}(s,a')\right)$.

Similarly, when $Z>\exp\left(-\eta_s\min_{a'\in\mathcal{A}_s}l^{(t)}(s,a')\right)$, for any $a\in\mathcal{A}_s$, we have $\frac{\bar{\pi}^{(t)}(a\,|\,s)\exp\left(-\eta_s l^{(t)}(s,a)\right)}{Z}<\bar{\pi}^{(t)}(a\,|\,s)$. It implies that $\sum_{a\in\mathcal{A}_s}\pi^{(t)}(a\,|\,s)<\sum_{a\in\mathcal{A}_s}\bar{\pi}^{(t)}(a\,|\,s)$, unless $\bar{\pi}^{(t)}(a\,|\,s)=\gamma_s\nu_{s,a}$ for all $a\in\mathcal{A}_s$, which is impossible since we assume $\gamma_s<1$. Therefore, $Z\le\exp\left(-\eta_s\min_{a'\in\mathcal{A}_s}l^{(t)}(s,a')\right)$.

Then, if $\frac{\bar{\pi}^{(t)}(a\,|\,s)\exp\left(-\eta_s l^{(t)}(s,a)\right)}{Z} \geq \gamma_s \nu_{s,a}$, we have

$$
\begin{aligned}
1 \leq & \frac{\pi^{(t)}(a\,|\,s)}{\bar{\pi}^{(t)}(a\,|\,s)} \\
= & \frac{\exp\left(-\eta_s l^{(t)}(s,a)\right)}{Z} \\
\leq & \exp\left(\eta_s \max_{a' \in \mathcal{A}_s} l^{(t)}(s,a') - \eta_s l^{(t)}(s,a)\right) \\
\leq & \exp\left(\frac{\eta_s}{\alpha_s\left(1+\eta_s\tau_s^{(t-1)}\right)}\left(\max_{a' \in \mathcal{A}_s} q^{(t)}(s,a') - q^{(t)}(s,a)\right) + \frac{\eta_s \tau_s^{(t-1)}}{1+\eta_s\tau_s^{(t-1)}} \log \max_{a' \in \mathcal{A}_s} \frac{\bar{\pi}^{(t)}(a'\,|\,s)}{\bar{\pi}^{(t)}(a\,|\,s)}\right) \\
\overset{(i)}{\leq} & \exp\left(\frac{\eta_s}{\alpha_s}\left(2\left\|\boldsymbol{q}\right\|_\infty + \frac{\tau\alpha_s}{M_1}\log\frac{1}{\gamma}\right)\right) \\
\overset{(ii)}{\leq} & 1 + 2\frac{\eta_s}{\alpha_s}\left(2\left\|\boldsymbol{q}\right\|_\infty + \frac{\tau\alpha_s}{M_1}\log\frac{1}{\gamma}\right)
\end{aligned}
$$

In (i) we use the fact $\left\|\boldsymbol{q}\right\|_\infty \geq \left|q^{(t)}(s,a)\right|$ for any $a \in \mathcal{A}_s$. In (ii) we use $e^x \leq 1+2x$ for $x \in [0,1]$.

If $\frac{\bar{\pi}^{(t)}(a\,|\,s)\exp\left(-\eta_s l^{(t)}(s,a)\right)}{Z} < \bar{\pi}^{(t)}(a\,|\,s)$, we have

$$
\begin{aligned}
\frac{\exp\left(-\eta_s l^{(t)}(s,a)\right)}{Z} \geq & \exp\left(\eta_s \min_{a' \in \mathcal{A}_s} l^{(t)}(s,a') - \eta_s l^{(t)}(s,a)\right) \\
\geq & \exp\left(-\frac{\eta_s}{\alpha_s}\left(2\left\|\boldsymbol{q}\right\|_\infty + \frac{\tau\alpha_s}{M_1}\log\frac{1}{\gamma}\right)\right).
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
1 \geq \frac{\pi^{(t)}(a\,|\,s)}{\bar{\pi}^{(t)}(a\,|\,s)} \geq \frac{\bar{\pi}^{(t)}(a\,|\,s)\exp\left(-\eta_s l^{(t)}(s,a)\right)}{\bar{\pi}^{(t)}(a\,|\,s)Z} \geq & \exp\left(-\frac{\eta_s}{\alpha_s}\left(2\left\|\boldsymbol{q}\right\|_\infty + \frac{\tau\alpha_s}{M_1}\log\frac{1}{\gamma}\right)\right) \\
\geq & 1 - \frac{\eta_s}{\alpha_s}\left(2\left\|\boldsymbol{q}\right\|_\infty + \frac{\tau\alpha_s}{M_1}\log\frac{1}{\gamma}\right).
\end{aligned}
$$

Then, for any $a \in \mathcal{A}_s$, we have

$$
\left|\frac{\pi^{(t)}(a\,|\,s)}{\bar{\pi}^{(t)}(a\,|\,s)} - 1\right| \leq 2\frac{\eta_s}{\alpha_s}\left(2\left\|\boldsymbol{q}\right\|_\infty + \frac{\tau\alpha_s}{M_1}\log\frac{1}{\gamma}\right).
$$

Therefore,

$$
\left\|\pi_{p(s)}^{(t)}(\cdot\,|\,s) - \bar{\pi}_{p(s)}^{(t)}(\cdot\,|\,s)\right\|_1 = \sum_{a \in \mathcal{A}_s} \bar{\pi}^{(t)}(a\,|\,s)\left|\frac{\pi^{(t)}(a\,|\,s)}{\bar{\pi}^{(t)}(a\,|\,s)} - 1\right| \leq 2\frac{\eta_s}{\alpha_s}\left(2\left\|\boldsymbol{q}\right\|_\infty + \frac{\tau\alpha_s}{M_1}\log\frac{1}{\gamma}\right).
$$

Similarly, the upperbound above also holds for $\left|\frac{\bar{\pi}^{(t+1)}(a\,|\,s)}{\bar{\pi}^{(t)}(a\,|\,s)} - 1\right|$. □

C.5.4 Update Rule of MWU

For ease of representation, we ignore the learning rate η without loss of generality (the update-rule is the same without η when multiplying both the gradient $\boldsymbol{g}$ and τ by η).

Lemma C.5.4. Consider the update-rule $\boldsymbol{x}^{(2)} = \operatorname{argmin}_{\boldsymbol{x}\in\Delta_{|\mathcal{A}|}^{\gamma,\boldsymbol{\nu}}} \langle \boldsymbol{x}, \boldsymbol{g}\rangle + \tau\psi(\boldsymbol{x}) + D_\psi(\boldsymbol{x}, \boldsymbol{x}^{(1)})$ where $\boldsymbol{x}^{(1)} \in \Delta_{|\mathcal{A}|}^{\gamma,\boldsymbol{\nu}}$, $\gamma \geq 0$ is a constant, and $\boldsymbol{\nu} \in \Delta^{|\mathcal{A}|}$, $\mathcal{A}$ is the action set and $\psi(\boldsymbol{x}) = \sum_{a\in\mathcal{A}} x\log x + \log|\mathcal{A}|$. Then, the update-rule is equivalent to

$$x_a^{(2)} = \max\left\{\frac{\left(x_a^{(1)}\right)^{\frac{1}{1+\tau}}\exp(-\frac{g_a}{1+\tau})}{Z}, \gamma\nu_a\right\} \tag{C.5.4}$$

for any $a \in \mathcal{A}$. $Z > 0$ is the normalizing constant to ensure $\sum_{a\in\mathcal{A}} x_a^{(2)} = 1$.

Proof. By definition of Bregman divergence, $D_\psi(\boldsymbol{x}, \boldsymbol{x}^{(1)}) = \sum_{a\in\mathcal{A}} x_a \log\frac{x_a}{x_a^{(1)}}$. Therefore, the Lagrangian function of the update-rule is

$$\mathcal{F}(\boldsymbol{x}, \alpha, \boldsymbol{\beta}) \coloneqq \langle \boldsymbol{x}, \boldsymbol{g}\rangle + \tau\sum_{a\in\mathcal{A}} x_a\log x_a + \sum_{a\in\mathcal{A}} x_a\log\frac{x_a}{x_a^{(1)}} + \alpha(\sum_{a\in\mathcal{A}} x_a - 1) + \sum_{a\in\mathcal{A}}\beta_a(x_a - \gamma\nu_a).$$

By taking $\nabla_{\boldsymbol{x}}\mathcal{F}(\boldsymbol{x}, \alpha, \boldsymbol{\lambda}) = 0$, for any $a \in \mathcal{A}$, we have

$$g_a + \tau\log x_a^{(2)} + \log\frac{x_a^{(2)}}{x_a^{(1)}} + 1 + \tau + \alpha + \beta_a = 0,$$

which implies that

$$x_a^{(2)} = \left(x_a^{(1)}\right)^{\frac{1}{1+\tau}}\exp\left(-\frac{1}{1+\tau}\left(g_a + 1 + \tau + \alpha + \beta_a\right)\right).$$

By duality, $\beta_a \leq 0$. By complementary slackness, we have $\beta_a(x_a^{(2)} - \gamma\nu_a) = 0$. Therefore, when $\beta_a < 0$, we have $x_a^{(2)} = \gamma\nu_a$, which implies that $\left(x_a^{(1)}\right)^{\frac{1}{1+\tau}}\exp\left(-\frac{1}{1+\tau}\left(g_a + 1 + \tau + \alpha\right)\right) < \gamma\nu_a$. When $\beta_a = 0$, we have $x_a^{(2)} \geq \gamma\nu_a$ so that $\left(x_a^{(1)}\right)^{\frac{1}{1+\tau}}\exp\left(-\frac{1}{1+\tau}\left(g_a + 1 + \tau + \alpha\right)\right) \geq \gamma\nu_a$. Therefore, the effect of β_a is equivalent to take a max on $\left(x_a^{(1)}\right)^{\frac{1}{1+\tau}}\exp\left(-\frac{1}{1+\tau}\left(g_a + 1 + \tau + \alpha\right)\right)$ and we have the update-rule

$$\begin{aligned} x_a^{(2)} &= \max\left\{\left(x_a^{(1)}\right)^{\frac{1}{1+\tau}}\exp\left(-\frac{1}{1+\tau}\left(g_a + 1 + \tau + \alpha\right)\right), \gamma\nu_a\right\} \\ &= \max\left\{\frac{\left(x_a^{(1)}\right)^{\frac{1}{1+\tau}}\exp(-\frac{g_a}{1+\tau})}{Z}, \gamma\nu_a\right\} \end{aligned}$$

where $Z = \exp\left(\frac{1+\tau+\alpha}{1+\tau}\right)$. □

Remark C.5.5. In practice, we can implement the update-rule in Lemma C.5.4 as follows. We assume $\gamma < 1$ since otherwise we can simply let $\boldsymbol{x}^{(2)} = \boldsymbol{\nu}$.

- Compute $\widehat{x}_a = \left(x_a^{(1)}\right)^{\frac{1}{1+\tau}} \exp(-\frac{g_a}{1+\tau})$ and sort it in increasing order, which is $\widehat{x}_1 \leq \widehat{x}_2 \leq \cdots \leq \widehat{x}_{|\mathcal{A}|}$. Simultaneously, adjusting $\boldsymbol{\nu}$ according to the sorting of $\widehat{\boldsymbol{x}}$ to get $\widehat{\boldsymbol{\nu}}$, which is the lowerbound $\widehat{\boldsymbol{x}}$ should satisfy.
- Enumerate $i = 0, 1, 2, \cdots, |\mathcal{A}|$. Let $Z = \frac{\sum_{j>i} \widehat{x}_j}{1-\gamma \sum_{j=1}^{i} \widehat{\nu}_j}$.
- Check $\widehat{x}_i \leq \gamma \widehat{\nu}_i Z$ if $i > 0$ and $\widehat{x}_{i+1} \geq \gamma \widehat{\nu}_{i+1} Z$ if $i < |\mathcal{A}|$. If the current Z satisfies, return it. Otherwise, continue the enumeration.

According to the monotonicity of $\max\left\{\frac{\left(x_a^{(1)}\right)^{\frac{1}{1+\tau}} \exp(-\frac{g_a}{1+\tau})}{Z}, \gamma\nu_a\right\}$ with respect to Z, the algorithm above will definitely find the correct Z and the time complexity is $\mathcal{O}(|\mathcal{A}| \log |\mathcal{A}|)$ (the bottleneck is the sort).

References

Kenshi Abe, Kaito Ariu, Mitsuki Sakamoto, and Atsushi Iwasaki. Adaptively perturbed mirror descent for learning in games. *International Conference on Machine Learning (ICML)*, 2024.

Yu Bai, Chi Jin, Song Mei, and Tiancheng Yu. Near-optimal learning of extensive-form games with imperfect information. In *International Conference on Machine Learning (ICML)*, 2022.

James P. Bailey and Georgios Piliouras. Multiplicative weights update in zero-sum games. In *ACM Conference on Economics and Computation (EC)*, 2018.

Anton Bakhtin, David J Wu, Adam Lerer, Jonathan Gray, Athul Paul Jacob, Gabriele Farina, Alexander H Miller, and Noam Brown. Mastering the game of no-press diplomacy via human-regularized reinforcement learning and planning. *International Conference on Learning Representations (ICLR)*, 2022.

David Balduzzi, Marta Garnelo, Yoram Bachrach, Wojciech Czarnecki, Julien Perolat, Max Jaderberg, and Thore Graepel. Open-ended learning in symmetric zero-sum games. In *International Conference on Machine Learning (ICML)*, 2019.

Christopher Berner, Greg Brockman, Brooke Chan, Vicki Cheung, Przemysław Dębiak, Christy Dennison, David Farhi, Quirin Fischer, Shariq Hashme, Chris Hesse, et al. Dota 2 with large scale deep reinforcement learning. *arXiv preprint arXiv:1912.06680*, 2019.

Michael Bowling, Neil Burch, Michael Johanson, and Oskari Tammelin. Heads-up Limit Hold'em poker is solved. *Science*, 347(6218):145–149, 2015.

Noam Brown and Tuomas Sandholm. Solving imperfect-information games via discounted regret minimization. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2019a.

Noam Brown and Tuomas Sandholm. Superhuman ai for multiplayer poker. *Science*, 365 (6456):885–890, 2019b.

Noam Brown, Adam Lerer, Sam Gross, and Tuomas Sandholm. Deep counterfactual regret minimization. In *International Conference on Machine Learning (ICML)*, 2019.

Yang Cai, Argyris Oikonomou, and Weiqiang Zheng. Tight last-iterate convergence of the extragradient and the optimistic gradient descent-ascent algorithm for constrained monotone variational inequalities. *arXiv preprint arXiv:2204.09228*, 2022.

Shicong Cen, Chen Cheng, Yuxin Chen, Yuting Wei, and Yuejie Chi. Fast global convergence of natural policy gradient methods with entropy regularization. *Operations Research*, 2021a.

Shicong Cen, Yuting Wei, and Yuejie Chi. Fast policy extragradient methods for competitive games with entropy regularization. In *Neural Information Processing Systems (NeurIPS)*, 2021b.

Constantinos Daskalakis and Ioannis Panageas. Last-iterate convergence: Zero-sum games and constrained min-max optimization. In *Innovations in Theoretical Computer Science (ITCS)*, 2019.

Constantinos Daskalakis, Andrew Ilyas, Vasilis Syrgkanis, and Haoyang Zeng. Training GANs with optimism. In *International Conference on Learning Representations (ICLR)*, 2018.

Gabriele Farina and Tuomas Sandholm. Model-free online learning in unknown sequential decision making problems and games. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2021.

Gabriele Farina, Christian Kroer, Noam Brown, and Tuomas Sandholm. Stable-predictive optimistic counterfactual regret minimization. In *International Conference on Machine Learning (ICML)*, 2019a.

Gabriele Farina, Christian Kroer, and Tuomas Sandholm. Online convex optimization for sequential decision processes and extensive-form games. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2019b.

Gabriele Farina, Christian Kroer, and Tuomas Sandholm. Optimistic regret minimization for extensive-form games via dilated distance-generating functions. *Neural Information Processing Systems (NeurIPS)*, 2019c.

Gabriele Farina, Christian Kroer, and Tuomas Sandholm. Stochastic regret minimization in extensive-form games. In *International Conference on Machine Learning (ICML)*, 2020.

Gabriele Farina, Christian Kroer, and Tuomas Sandholm. Faster game solving via predictive blackwell approachability: Connecting regret matching and mirror descent. In *AAAI Conference on Artificial Intelligence*, 2021a.

Gabriele Farina, Robin Schmucker, and Tuomas Sandholm. Bandit linear optimization for sequential decision making and extensive-form games. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2021b.

Côme Fiegel, Pierre Ménard, Tadashi Kozuno, Rémi Munos, Vianney Perchet, and Michal Valko. Adapting to game trees in zero-sum imperfect information games. In *International Conference on Machine Learning (ICML)*, 2023.

Côme Fiegel, Pierre Ménard, Tadashi Kozuno, Rémi Munos, Vianney Perchet, and Michal Valko. Local and adaptive mirror descents in extensive-form games. *Neural Information Processing Systems (NeurIPS)*, 2024.

Haobo Fu, Weiming Liu, Shuang Wu, Yijia Wang, Tao Yang, Kai Li, Junliang Xing, Bin Li, Bo Ma, Qiang Fu, et al. Actor-critic policy optimization in a large-scale imperfect-information game. In *International Conference on Learning Representations (ICLR)*, 2021.

Matthieu Geist, Bruno Scherrer, and Olivier Pietquin. A theory of regularized markov decision processes. In *International Conference on Machine Learning (ICML)*, 2019.

Andrew Gilpin, Javier Pena, and Tuomas W Sandholm. First-order algorithm with o (ln $(1/\varepsilon)$) convergence for-equilibrium in two-person zero-sum games. 2008.

Audrūnas Gruslys, Marc Lanctot, Rémi Munos, Finbarr Timbers, Martin Schmid, Julien Perolat, Dustin Morrill, Vinicius Zambaldi, Jean-Baptiste Lespiau, John Schultz, et al. The advantage regret-matching actor-critic. *arXiv preprint arXiv:2008.12234*, 2020.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International Conference on Machine Learning (ICML)*, 2018.

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

Samid Hoda, Andrew Gilpin, Javier Peña, and Tuomas Sandholm. Smoothing techniques for computing Nash Equilibria of sequential games. *Math. Oper. Res.*, 35(2):494–512, 2010. doi:10.1287/moor.1100.0452.

Josef Hofbauer and Ed Hopkins. Learning in perturbed asymmetric games. *Games and Economic Behavior*, 52(1):133–152, 2005.

Shengyi Huang, Rousslan Fernand Julien Dossa, Chang Ye, Jeff Braga, Dipam Chakraborty, Kinal Mehta, and João G.M. Araújo. Cleanrl: High-quality single-file implementations of deep reinforcement learning algorithms. *Journal of Machine Learning Research*, 23(274): 1–18, 2022. URL http://jmlr.org/papers/v23/21-1342.html.

Julian Ibarz, Jie Tan, Chelsea Finn, Mrinal Kalakrishnan, Peter Pastor, and Sergey Levine. How to train your robot with deep reinforcement learning: lessons we have learned. *The International Journal of Robotics Research*, 40(4-5):698–721, 2021.

B Ravi Kiran, Ibrahim Sobh, Victor Talpaert, Patrick Mannion, Ahmad A Al Sallab, Senthil Yogamani, and Patrick Pérez. Deep reinforcement learning for autonomous driving: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 23(6):4909–4926, 2021.

Tadashi Kozuno, Pierre Ménard, Remi Munos, and Michal Valko. Learning in two-player zero-sum partially observable markov games with perfect recall. *Neural Information Processing Systems (NeurIPS)*, 2021.

Christian Kroer, Kevin Waugh, Fatma Kılınç-Karzan, and Tuomas Sandholm. Faster algorithms for extensive-form game solving via improved smoothing functions. *Mathematical Programming*, 179(1):385–417, 2020.

Harold W Kuhn. A simplified two-person poker. *Contributions to the Theory of Games*, 1 (417):97–103, 1950.

Marc Lanctot, Kevin Waugh, Martin Zinkevich, and Michael Bowling. Monte carlo sampling for regret minimization in extensive games. *Neural Information Processing Systems (NeurIPS)*, 2009.

Marc Lanctot, Edward Lockhart, Jean-Baptiste Lespiau, Vinicius Zambaldi, Satyaki Upadhyay, Julien Pérolat, Sriram Srinivasan, Finbarr Timbers, Karl Tuyls, Shayegan Omidshafiei, et al. Openspiel: A framework for reinforcement learning in games. *arXiv preprint arXiv:1908.09453*, 2019.

Chung-Wei Lee, Christian Kroer, and Haipeng Luo. Last-iterate convergence in extensive-form games. In *Neural Information Processing Systems (NeurIPS)*, 2021.

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 (AISTATS)*, 2021.

Stefanos Leonardos, Georgios Piliouras, and Kelly Spendlove. Exploration-exploitation in multi-agent competition: convergence with bounded rationality. *Neural Information Processing Systems (NeurIPS)*, 2021.

Tengyuan Liang and James Stokes. Interaction matters: A note on non-asymptotic local convergence of generative adversarial networks. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2019.

Mingyang Liu, Asuman E. Ozdaglar, Tiancheng Yu, and Kaiqing Zhang. The power of regularization in solving extensive-form games. In *International Conference on Learning Representations (ICLR)*, 2023.

Mingyang Liu, Gabriele Farina, and Asuman Ozdaglar. LiteEFG: An efficient python library for solving extensive-form games, 2024a. URL https://arxiv.org/abs/2407.20351.

Mingyang Liu, Gabriele Farina, and Asuman Ozdaglar. A policy-gradient approach to solving imperfect-information games with iterate convergence. *arXiv preprint arXiv:2408.00751*, 2024b.

Weiming Liu, Huacong Jiang, Bin Li, and Houqiang Li. Equivalence analysis between counterfactual regret minimization and online mirror descent. In *International Conference on Machine Learning (ICML)*, 2022.

Stephen Marcus McAleer, Gabriele Farina, Marc Lanctot, and Tuomas Sandholm. ESCHER: eschewing importance sampling in games by computing a history value function to estimate regret. In *International Conference on Learning Representations (ICLR)*, 2023.

Jincheng Mei, Chenjun Xiao, Csaba Szepesvári, and Dale Schuurmans. On the global convergence rates of softmax policy gradient methods. In *International Conference on Machine Learning (ICML)*, 2020.

Panayotis Mertikopoulos, Christos H. Papadimitriou, and Georgios Piliouras. Cycles in adversarial regularized learning. In *Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2018.

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 (ICLR)*, 2019.

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

Aryan Mokhtari, Asuman E. 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 (AISTATS)*, 2020.

Matej Moravčík, Martin Schmid, Neil Burch, Viliam Lisỳ, Dustin Morrill, Nolan Bard, Trevor Davis, Kevin Waugh, Michael Johanson, and Michael Bowling. Deepstack: Expert-level artificial intelligence in heads-up no-limit poker. *Science*, 356(6337):508–513, 2017.

Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2003.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Neural Information Processing Systems (NeurIPS)*, 2022.

Julien Perolat, Remi Munos, Jean-Baptiste Lespiau, Shayegan Omidshafiei, Mark Rowland, Pedro Ortega, Neil Burch, Thomas Anthony, David Balduzzi, Bart De Vylder, et al. From poincaré recurrence to convergence in imperfect information games: Finding equilibrium via regularization. In *International Conference on Machine Learning (ICML)*, 2021.

Julien Perolat, Bart De Vylder, Daniel Hennes, Eugene Tarassov, Florian Strub, Vincent de Boer, Paul Muller, Jerome T Connor, Neil Burch, Thomas Anthony, et al. Mastering the game of stratego with model-free multiagent reinforcement learning. *Science*, 378(6623): 990–996, 2022.

Georgios Piliouras, Lillian Ratliff, Ryann Sim, and Stratis Skoulakis. Fast convergence of optimistic gradient ascent in network zero-sum extensive form games. *arXiv preprint arXiv:2207.08426*, 2022.

Alexander Rakhlin and Karthik Sridharan. Optimization, learning, and games with predictable sequences. In *Neural Information Processing Systems (NeurIPS)*, 2013.

J Ben Rosen. Existence and uniqueness of equilibrium points for concave n-person games. *Econometrica: Journal of the Econometric Society*, pages 520–534, 1965.

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

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *nature*, 550(7676):354–359, 2017.

Samuel Sokota, Ryan D'Orazio, J. Zico Kolter, Nicolas Loizou, Marc Lanctot, Ioannis Mitliagkas, Noam Brown, and Christian Kroer. A unified approach to reinforcement learning, quantal response equilibria, and two-player zero-sum games. In *International Conference on Learning Representations (ICLR)*, 2023.

Finnegan Southey, Michael Bowling, Bryce Larson, Carmelo Piccione, Neil Burch, Darse Billings, and Chris Rayner. Bayes' bluff: opponent modelling in poker. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2005.

Eric Steinberger. Single deep counterfactual regret minimization. *arXiv preprint arXiv:1901.07621*, 2019.

Eric Steinberger, Adam Lerer, and Noam Brown. Dream: Deep regret minimization with advantage baselines and model-free learning. *arXiv preprint arXiv:2006.10410*, 2020.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

Oskari Tammelin, Neil Burch, Michael Johanson, and Michael Bowling. Solving Heads-Up Limit Texas Hold'em. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2015.

Karl Tuyls, Katja Verbeeck, and Tom Lenaerts. A selection-mutation model for q-learning in multi-agent systems. In *Proceedings of the second international joint conference on Autonomous agents and multiagent systems*, pages 693–700, 2003.

Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, et al. Grandmaster level in starcraft ii using multi-agent reinforcement learning. *nature*, 575 (7782):350–354, 2019.

Bernhard Von Stengel. Efficient computation of behavior strategies. *Games and Economic Behavior*, 14(2):220–246, 1996.

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)*, 2021.

Martin Zinkevich, Michael Johanson, Michael H. Bowling, and Carmelo Piccione. Regret minimization in games with incomplete information. In *Neural Information Processing Systems (NeurIPS)*, 2007.