Provably Efficient Policy Optimization for Two-Player Zero-Sum Markov Games

Yulai Zhao
Tsinghua University

Yuandong Tian
Meta AI Research

Jason D. Lee
Princeton University

Simon S. Du
University of Washington
Meta AI Research

Abstract

Policy-based methods with function approximation are widely used for solving two-player zero-sum games with large state and/or action spaces. However, it remains elusive how to obtain optimization and statistical guarantees for such algorithms. We present a new policy optimization algorithm with function approximation and prove that under standard regularity conditions on the Markov game and the function approximation class, our algorithm finds a near-optimal policy within a polynomial number of samples and iterations. To our knowledge, this is the first provably efficient policy optimization algorithm with function approximation that solves two-player zero-sum Markov games.

1 Introduction

Two-player zero-sum Markov game is a popular setting with many applications, such as Go (Silver et al., 2016), StarCraft II (Vinyals et al., 2019), and poker (Brown and Sandholm, 2018). In this setting, the goal of player one is to find a policy that achieves the maximum reward against player two who plays optimally to minimize the reward in response to player one's policy.

Policy optimization methods are widely used for solving zero-sum games. These algorithms often constrain the policy in a parametric form, and compute the gradient of the cumulative reward with respect to the parameters using the policy gradient theorem or its variants to update the parameters iteratively (Sutton et al., 2000; Kakade, 2002; Silver et al., 2014). Due to its flexibility, a wide range of successful results are attained by policy optimization methods. For example, Lockhart et al. (2019) performed direct policy optimization against worst-case opponents and empirically demonstrate their effectiveness in Kuhn Poker and Goofspiel card game. Foerster et al. (2017) invented LOLA where each agent shapes the learning of other agents. It gave the highest average returns on the iterated prisoners' dilemma (IPD).

Despite the large body of empirical work using policy optimization methods for two-player zero-sum Markov games, theoretical studies are very limited. In this paper, we aim to answer the following fundamental question:

Can we design a provably efficient policy optimization algorithm with function approximation for two-player zero-sum Markov games with a large state-action space?

We answer the above question affirmatively. We summarize our contributions below.

Our contributions. We design a new, provably efficient policy optimization algorithm for two-player zero-sum Markov games based on the natural policy gradient (NPG) method (Kakade, 2002). On a high level, our algorithm has the two-step style as in previous work on value-based algorithms for two-player zero-sum Markov games (Perolat et al., 2015). In the **Greedy step**, we aim to find *a pair of policies* that approximately solves matrix games for a given value function, and in the **Iteration step**, we aim to update the value function upon the current policy. In contrast to their value-based algorithm where function approximation is used for value functions, our results are entirely policy-based. We only have *function approximation for policies*. Therefore, we need to tackle additional challenges which are absent in value-based algorithms.

1. First, in the **Greedy step**, the algorithms in (Perolat et al., 2015) requires solving two-

Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151.

player zero-sum matrix games *for every state* in a value-based manner. The computational complexity scales with the size of the state-action space, which can be infeasible. For finding the equilibria for state-wise matrix games, we employ policy-based methods whose sample complexity only scales with the complexity of the function class (e.g., feature dimension for linear function approximation) instead of the size of state-action space. Specifically, we design a subroutine that combines two-player policy gradients with the optimistic mirror descent (OMD) updates (Rakhlin and Sridharan, 2013) to solve the zero-sum matrix game in a computational and statistical efficient way.

2. Second, in the **Iteration step**, Perolat et al. (2015) used *Generalized Policy Iteration* to evaluate the value function while we only have function approximation for policies. We leverage recent developments on NPG in single-agent RL (Agarwal et al., 2020) to update *policies* in this step and represent the value function using *policies* instead of explicitly storing the value function.

3. Third, technically, we incorporates policy-based methods into value-based schemes, and develop new perturbation analyses for policy-based methods, both of which may be of independent interest.

Theoretically, first, to illustrate the main idea of our algorithm, in Section 4, we study an idealized "population" tabular Markov game setting where we can access the population quantities, including the true policy gradients and the Fisher information. We prove an $\widetilde{O}\left(\frac{1}{T}\right)$ rate,[1] where T is the number of iterations. This result is interesting in its own right because this matches the rate in the single-agent RL setting (Agarwal et al., 2020). We further obtain an improved rate in the entropy-regularized setting (Cen et al., 2020).

We present our main algorithm and theoretical results in Section 5 where we study Markov games with a large state-action space, and log-linear policy parameterization is used for generalization. Instead of the idealized "population" setting, we study the realistic online setting, where we can only access the model through interactions. We prove an $\widetilde{O}\left(\frac{1}{\sqrt{T}}+\frac{1}{N^{1/4}}\right)$ rate where T is the number of iterations and N is the number of samples (interaction with the model). To our knowledge, this is the first quantitative analysis of online policy optimization methods with function approximation for two-player zero-sum Markov games.

[1] $\widetilde{O}(\cdot)$ hides logarithmic factors.

2 Related Work

A large number of empirical works have proven the validity and efficiency of PG/NPG based methods in games and other applications (Silver et al., 2016, 2017; Guo et al., 2016; Mousavi et al., 2017; Tian et al., 2019). Below we mostly focus on relevant algorithmic and theoretical papers.

There is a long line of work developing computationally efficient algorithms for multi-agent RL in Markov games. Value-based approaches (Shapley, 1953; Patek, 1997; Littman, 1994; Bai and Jin, 2020; Bai et al., 2020) try to find the optimal value function. When the size of the state-action space is large, Approximate Dynamic Programming (ADP) techniques are often incorporated into value-based methods. Extending the error propagation scheme of ADP developed by Scherrer et al. (2012) to two-player zero-sum games, Perolat et al. (2015) obtained a performance bound in general norms. Using this error propagation scheme on ADP, Pérolat et al. (2016) adapted three value-based algorithms (PSDP, NSVI, NSPI) to the two-player zero-sum setting. Recently, Yu et al. (2019) replaced the policy evaluation step of Approximate Modified Policy Iteration (AMPI) introduced by Scherrer et al. (2015) with function approximation in a Reproducing Kernel Hilbert Space (RKHS) and proved linear convergence to l_∞-norm up to a statistical error. While focusing on policy-based methods, our paper also leverages the error propagation analysis (Perolat et al., 2015).

Another type of algorithms on two-player zero-sum Markov games is policy-based. One family of algorithms is based on fictitious play (Brown, 1951; Robinson, 1951). Fictitious play is a classical strategy proposed by Brown (1951), where each player adopts a policy that best responds to the average policy of other agents inferred from historical data. For example, Heinrich et al. (2015) introduced two variants of fictitious play: 1) an algorithm for extensive-form games which is realization-equivalent to its normal-form counterpart, 2) Fictitious Self-Play (FSP) which is a framework computing the best response via fitted Q-iteration. Our paper also aims to find the best response iteratively. Another family of policy-based methods is based on the idea of counterfactual regret minimization (CFR) (Zinkevich et al., 2008). Brown and Sandholm (2019) invented a novel CFR variant which utilizes techniques such as reweighting iterations and leveraging optimistic regret matching. Although these two families of algorithms are similar to ours in spirit, they are quite different technically and their theoretical analysis does not apply to our setting.

The current paper focuses on using NPG techniques for solving two-player zero-sum Markov games. NPG is first introduced by Kakade (2002) to better explore the underlying structure of the reinforcement learning (RL) problem instance. Extensions of NPG methods are also used to solve zero-sum games. Zhang et al. (2019); Bu et al. (2019) applied projected natural nested gradient under a linear quadratic setting, a significant class of zero-sum Markov games. Extensions to imitation learning were also studied in (Song et al., 2018).

In terms of theoretical analysis on PG/NPG methods, Agarwal et al. (2020) showed that tabular NPG could provide an $\mathcal{O}(1/T)$ iteration complexity, as well as a sample complexity of $\mathcal{O}(1/N^{\frac{1}{4}})$ for online NPG with function approximation. In contrast, we provide bounds for the two-player zero-sum case, which is significantly more challenging. In two-player zero-sum games, the non-stationary environment faced by each individual agent invalidates the stationary structure of the single-agent setting, and thus precludes the direct application of the convergence proof from the single-agent setting. Furthermore, each agent in two-player zero-sum games must adapt to the other agent's policy, which poses additional difficulties. Zhang et al. (2020) proposed a new variant of PG methods that yielded unbiased estimates of policy gradients, which enabled non-convex optimization tools to be applied in establishing global convergence. Despite being non-convex, Agarwal et al. (2020); Bhandari and Russo (2019) identified structural properties of finite Markov decision processes (MDPs): the objective function has no suboptimal local minimum. They further gave conditions under which any local minimum is near-optimal.

Schulman et al. (2015) developed a practical algorithm called TRPO which could be seen as a KL divergence-constrained variant of NPG. They show monotonic improvements of the expected return during optimization. Shani et al. (2020) considered a sample-based TRPO (Schulman et al., 2015) and proved an $\tilde{\mathcal{O}}(1/\sqrt{N})$ convergence rate to the global optimum, which could be improved to $\tilde{\mathcal{O}}(1/N)$ when regularized. Cen et al. (2020) showed that fast convergence rate of NPG methods can be obtained with entropy regularization. Applying NPG to linear quadratic games, Zhang et al. (2019) and Bu et al. (2019) proved that: for finding Nash equilibrium, NPG enjoys sublinear convergence rate. Both analyses rely on the linearity of the dynamics which does not hold in general Markov games considered in this paper.

Recently, Daskalakis et al. (2020) showed independent policy gradient methods converge to a min-max equilibrium. Compared to our work, they focused on the tabular case and did not study the function approximation. They also assumed that the probability of stopping at any state (Daskalakis et al., 2020, Section 2) is bounded below from a certain positive number, which is not a standard modelling approach and is hard to validate empirically. We instead use concentrability coefficients as a characterization of the game structure (cf. Definition 1). In general, these two conditions do not imply each other. Comparing with their work, ours is cheap in sample complexity. To find an ϵ-optimal solution, their sample complexity has an $O\left(\epsilon^{-12.5}\right)$ scaling whereas ours has an $O\left(\epsilon^{-6}\right)$ scaling.

Our work is related to Optimistic Mirror Descent (OMD) and its behavior in zero-sum games, which have received more attention lately. Daskalakis et al. (2018) proposed the use of optimistic mirror decent for training Wasserstein GANs to address the limit cycling problem in experiments. They also proved convergence to a equilibrium in bilinear zero-sum games. Generalizing (Daskalakis et al., 2018), Mertikopoulos et al. (2019) showed OMD converged in a class of non-monotone problems satisfying *coherence*. Their work made concrete steps toward establishing convergence beyond convex-concave games.

3 Preliminaries

In this section, we introduce the material background on two-player zero-sum Markov games and specify several quantities which will be used to analyze our algorithms for different settings.

3.1 Two-Player zero-sum Markov Games.

In this paper, we consider the centralized setting where we can control both players in the training phase to learn good policies. we focus on infinite-horizon discounted two-player zero-sum Markov games, which can be described by a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, r, \gamma)$: a set of states $\mathcal{S}$, a set of actions $\mathcal{A}$, a transition probability $\mathcal{P} : \mathcal{S} \times \mathcal{A} \times \mathcal{A} \to \Delta(\mathcal{S})$, a reward function $r : \mathcal{S} \times \mathcal{A} \times \mathcal{A} \to [0, 1]$, and a discount factor $\gamma \in [0, 1)$. We let σ to be the initial state distribution and define policies as probability distributions over the action space: $x, f \in \mathcal{S} \to \Delta(\mathcal{A})$. [2] The value function $V^{x,f} : \mathcal{S} \to \mathbb{R}$ is defined as:

$$V^{x,f}(s) = \mathop{\mathbb{E}}_{\substack{a_t \sim x(\cdot|s_t) \\ b_t \sim f(\cdot|s_t) \\ s_{t+1} \sim \mathcal{P}(\cdot|s_t, a_t, b_t)}} \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t, b_t) \big| s_0 = s \right].$$

[2] For clarity we assume two players share the same set of actions, it is straight forward to generalize to the setting where two action sets are different. See Section 5.

we use distribution σ as the optimization measure we use to train the policy and use distribution ρ as the performance measure of our interest. We remark that these two separate measures are widely used in analyzing approximate dynamic programming and policy gradient (Agarwal et al., 2020; Perolat et al., 2015). We overload notations and define $V^{x,f}(\rho)$ as the expected value function of interest, i.e. $V^{x,f}(\rho) \coloneqq \mathbb{E}_{s\sim\rho}V^{x,f}(s)$.

In a two-player zero-sum Markov game, player one (x) wants to maximize the value function and the other player (f) wants to minimize it. We define the Markov game's state-action value function $Q^{x,f}: \mathcal{S}\times\mathcal{A}\times\mathcal{A}\to\mathbb{R}$, the advantage function $A^{x,f}: \mathcal{S}\times\mathcal{A}\times\mathcal{A}\to\mathbb{R}$, and the state visitation function $d_{s_0}^{x,f}: \mathcal{S}\to[0,1]$ as

$$Q^{x,f}(s,a,b) = r(s,a,b) + \gamma \mathop{\mathbb{E}}_{s'\sim\mathcal{P}(\cdot|s,a,b)} V^{x,f}(s'),$$
$$A^{x,f}(s,a,b) = Q^{x,f}(s,a,b) - V^{x,f}(s),$$
$$d_{s_0}^{x,f}(s) = (1-\gamma)\sum_{t=0}^{\infty}\gamma^t Pr(s_t = s|s_0,x,f)$$

where $s_0\in\mathcal{S}$ is an initial state, respectively. With the state visitation function at hand, we are prepared to introduce NPG (Kakade, 2002) for two-player zero-sum games which relies on the Fisher information matrix. Given player one's policy x, player two's policy f parameterized by θ, and starting state distribution σ, we define the Fisher information matrix $F_\sigma(\theta)$ as:

$$F_\sigma(\theta) = \mathbb{E}_{s\sim d_\sigma^{x,f}}\mathbb{E}_{b\sim f(\cdot|s)}\nabla_\theta \log f(b|s)\nabla_\theta \log f(b|s)^\top,$$

where we denote $d_\sigma^{x,f} = \mathbb{E}_{s_0\sim\sigma}d_{s_0}^{x,f}$ by the expectation form of the state visitation distribution.

One important concept in RL is the Bellman operator. For two-player zero-sum Markov games and two behavior policies x and f, we define $\mathcal{P}_{x,f}(s'|s) = \mathbb{E}_{a\sim x(\cdot|s),b\sim f(\cdot|s)}\mathcal{P}(s'|s,a,b)$ which performs as the transition kernel from s to any $s'\in\mathcal{S}$ and $r_{x,f}(s) = \mathbb{E}_{a\sim x(\cdot|s),b\sim f(\cdot|s)}r(s,a,b)$ which represents the reward each player can expect with policies (x,f). Bellman operators $\mathcal{T}_{x,f}, \mathcal{T}_x, \mathcal{T}$ act on any value function $v:\mathcal{S}\to\mathbb{R}$ and update it

- $\mathcal{T}_{x,f}v \coloneqq r_{x,f} + \gamma\mathcal{P}_{x,f}v$, which generalizes the standard Bellman operator.
- $\mathcal{T}_x v \coloneqq \inf_f \mathcal{T}_{x,f}v$, which is an asymmetric operator by letting f to be optimal. [3]
- $\mathcal{T}v \coloneqq \sup_x \mathcal{T}_x v = \sup_x \inf_f \mathcal{T}_{x,f}v$, which generalizes the standard Bellman optimality operator. It reflects notions of minimax equilibrium in essence.

[3] Here we focus on max player x (see Eq. 3). If we replace x by f, we will have an analogous notation for min player.

Perolat et al. (2015) introduced these operators as generalized counterparts of single agent RL. We are able to adopt the dynamic programming scheme only once the Bellman operators are introduced.

Since we are considering a learning problem, we need to collect samples from the environment. We assume we can stop and restart at any time. With this, we can have the following sampling oracle.

Episodic Sampling Oracle For a fixed state-action distribution ν_0, we can start from $s_0, a_0, b_0 \sim \nu_0$, act according to any policy pair (x,f), and terminate when desired. We obtain unbiased estimates of the *on policy* state-action distribution

$$\nu_{\nu_0}^{x,f}(s,a,b) = (1-\gamma)\cdot$$
$$\mathbb{E}_{s_0,a_0,b_0\sim\nu_0}\sum_{t=0}^{\infty}\gamma^t \Pr(s_t = s, a_t = a, b_t = b|s_0,a_0,b_0) \tag{1}$$

which can be used for acquiring an unbiased $Q^{x,f}(s,a,b)$ where $s,a,b\sim\nu_{\nu_0}^{x,f}$. See (Agarwal et al., 2020, Algorithm 1) for a sampler.

This oracle essentially requires that we can terminate at any time and restart, therefore many real-world applications including games and physics simulation (e.g., OpenFOAM (Weller et al., 1998)) admit this oracle. This oracle is also used in the analysis (Agarwal et al., 2020) (see Algorithm 1,3 and Assumption 6.3 therein).

The oracle is essential in analysis, technically because policy gradient methods need to estimate the values. This is the same reason as in the single-agent setting (Agarwal et al., 2020). Moreover, we believe this sampling oracle is not a strong assumption: it only requires that we can terminate at any time and restart. This is much weaker than the generative model assumption. The generative model assumes that one can query *any* state-action pair where we only require we can restart from a fixed initial distribution.

Shapley (1953) show that (x^*, f^*) is a pair of Nash equilibrium (NE) if the following inequalities hold for any state distribution ρ and policy pair (x,f):

$$V^{x,f^*}(\rho) \le V^{x^*,f^*}(\rho) = V^*(\rho) \le V^{x^*,f}(\rho). \tag{2}$$

NE always exists for discounted two-player zero-sum Markov Games (Filar and Vrieze, 2012). In practice, we seek to find an approximate pair of NE instead of an exact solution. The goal of this paper is to output a policy x that makes the metric

$$V^*(\rho) - \inf_f V^{x,f}(\rho) \tag{3}$$

small where ρ is some state distribution of interest. This metric measures the performance

of x against the worst-case f. If it is less than ϵ, we call x an one-sided ϵ-approximate NE, [4] it has been used by (Daskalakis et al., 2007; Göös and Rubinstein, 2018; Deligkas et al., 2017; Babichenko and Rubinstein, 2020; Daskalakis et al., 2020).

3.2 Function Approximation

This paper studies function approximation to generalize across a large state space in Section 5. To represent both behavior policies x and f, we adopt a log-linear parameterization: for a coefficient vector $\theta \in \mathbb{R}^d$, the associated probability of choosing action a under state s, $\pi_\theta(a|s)$, is given by $\frac{\exp(\theta^\top \phi_{s,a})}{\sum_{a' \in \mathcal{A}} \exp(\theta^\top \phi_{s,a'})}$ where $\phi_{s,a}$ is a feature vector representation of s and a. This parameterization has been used in (Branavan et al., 2009; Gimpel and Smith, 2010; Heess et al., 2013). We impose a regularity condition such that every $\|\phi_{s,a}\|_2 \leq D$. Note that log-linear parameterization is D^2-smooth in terms of θ (Agarwal et al., 2020). [5] This term is also known as *Policy Smoothness* when analyzing PG methods.

3.3 Problem-Dependent Quantities.

Our analysis relies on several problem-dependent quantities. We denote weighted L_p-norm of function f on state space $\mathcal{S}$ as $\|f\|_{p,\rho} = \left(\sum_{s \in \mathcal{S}} \rho(s)|f(s)|^p\right)^{\frac{1}{p}}$.

The first problem-dependent quantity is used to measure the inherent dynamics of Markov games.

Definition 1 (Concentrability Coefficients). *Given two distributions over states: ρ and σ. When σ is element-wise positive, define*

$$c_{\rho,\sigma}(j) = \sup_{x^1, f^1, \cdots x^j, f^j \in \mathcal{S} \to \Delta(\mathcal{A})} \left\| \frac{\rho \mathcal{P}_{x^1,f^1} \cdots \mathcal{P}_{x^j,f^j}}{\sigma} \right\|_\infty,$$

$$\mathcal{C}'_{\rho,\sigma} = (1-\gamma)^2 \sum_{m \geq 1} m\gamma^{m-1} c_{\rho,\sigma}(m-1),$$

$$\mathcal{C}^{l,k,d}_{\rho,\sigma} = \frac{(1-\gamma)^2}{\gamma^l - \gamma^k} \sum_{i=l}^{k-1} \sum_{j=i}^{\infty} \gamma^j c_{\rho,\sigma}(j+d).$$

Here, $x^1, f^1, \cdots x^j, f^j$ are j pairs of policies. Intuitively, the first term quantifies the distribution shift after taking j pairs of steps starting from ρ. The second term describes the accumulative effect of discounted distribution shifts. Finally, the last term represents the additive performance of $(k-l)$ accumulative distribution shift and thus it is often considered as stricter condition. See (Scherrer, 2014) for a thorough comparison on these coefficients. Generally speaking, if σ is sufficiently diverse across states, then these quantities are bounded from above. (Chen and Jiang, 2019) pointed out that small concentrability coefficients reflect a restriction on the MDPs dynamics. Concentrability coefficients are widely used in analyzing the convergence of approximate dynamic programming algorithms (Munos, 2005; Antos et al., 2008; Scherrer, 2014; Perolat et al., 2015) and recently in analyzing PG methods (Agarwal et al., 2020). In particular, (Agarwal et al., 2020) gave an example to show the *dependency on concentrability coefficients is necessary*. In these papers, their upper bounds all depend on the concentrability coefficients. For our two-player setting, we use the same definition of concentrability coefficients as (Perolat et al., 2015).

The second quantity measures how well a parameterized class can approximate in terms of a metric.

Definition 2 (Approximation Error). *Given a space $\mathcal{W}$ and a loss function $L : \mathcal{W} \to R$, we define $\epsilon_{approx} = \min_{w \in \mathcal{W}} L(w)$ as the approximation error of $\mathcal{W}$.*

This concept is widely used for analyzing function approximation (Menache et al., 2005; Jiang et al., 2015), state abstractions schemes (Jiang et al., 2015) and representation learning in RL (Bellemare et al., 2019). It explicitly describes the capacity of a parameter set.

4 Warm-up: Population Algorithm for Tabular Case

We first introduce the population version algorithm for the tabular case with the exact Fisher information matrix and policy gradients. The algorithm is spiritually similar to fictitious play. We enforce x, f to be tabular softmax parameterized by $\xi, \theta \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{A}|}$

Parameterization For vector $\theta \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{A}|}$, the probability associates to choosing action a under state s, $\pi_\theta(a|s)$, equals $\frac{\exp(\theta_{s,a})}{\sum_{a' \in \mathcal{A}} \exp(\theta_{s,a'})}$. One can verify that π_θ is 1-smooth in terms of θ.

This algorithm can be viewed as a prototypical algorithm and in the subsequent section, we will generalize to the online setting. The pseudo-code is listed in Algorithm 1.

In Algorithm 1, we perform K outer loops and obtain a near-optimal x and value function V_K. We note that this algorithm is asymmetric since our metric (Eq. 3)

[4] From an optimization perspective, the sampling complexity of finding a solution so that both the min and max player are approximate NE scales only twice as large as that in one-sided case, since we may apply algorithms with the roles switched.

[5] We follow standard smoothness definition. A function f is said to be β-smooth if for all $x, x' \in \mathbb{R}^d$: $\|\nabla f(x) - \nabla f(x')\|_2 \leq \beta \|x - x'\|_2$.

Algorithm 1 Population Two-Player NPG.

Input: $V_0 = 0$ a value function.
Output: Approximate policy x^K at Nash equilibrium
for $k = 1, 2, \cdots, K$ **do**
Greedy Step:
Run Algorithm 2 with A_s defined in Eq. 4 and returns $x^k(\cdot|s)$ for every state s.
Iteration Step:
Fix $x = x^k$, initialize $\theta = 0$.
for $t = 0, 1, \cdots, T-1$ **do**
$\theta^{t+1} = \theta^t - \eta F_\sigma(\theta^t)^\dagger \nabla_\theta V^{x, f^t}(\sigma)$.
end for
$V_k = V^{x, f^T}$.
end for

Algorithm 2 Subroutine: OMD for tabular case

Input: $f_0, g'_0, x_0, y'_0 \in \mathrm{Unif}(\mathcal{A})$, $\beta = \frac{1}{T'^2}$, and A_s for $s \in \mathcal{S}$.
Output: Approximate optimal $x_{\bar{T}'}$ for max player
for $t = 1, 2, \cdots, T'$ **do**
min player:
play $f_t(\cdot|s)$, observe $A_s^\top x_t(\cdot|s)$. Update:

$$g_t(i) \propto g'_{t-1}(i) e^{-\eta_t [x_t^\top A]_i}, \; g'_t = (1-\beta) g_t + \frac{\beta}{|\mathcal{A}|}\mathbf{I},$$

$$f_{t+1}(i) \propto g'_t(i) e^{-\eta_{t+1} [x_t^\top A]_i}$$

max player:
play $x_t(\cdot|s)$, observe $A_s f_t(\cdot|s)$. Update:

$$y_t(i) \propto y'_{t-1}(i) e^{-\eta'_t [A f_t]_i}, \; y'_t = (1-\beta) y_t + \frac{\beta}{|\mathcal{A}|}\mathbf{I},$$

$$x_{t+1}(i) \propto y'_t(i) e^{-\eta'_{t+1} [A f_t]_i}$$

end for

is only considering max player x while taking the best response of min player f.

Each outer iteration begins with a **Greedy Step**. For current V_{k-1}, we aim to find approximate equilibrium (x, f) with which $\mathcal{T}_{x,f} V \approx \mathcal{T} V_{k-1}$. This step is spiritually equivalent to finding minimax equilibrium of a matrix game for *every state* s. In intuition, this step helps to update V_{k-1} towards V^* (cf. contraction Lemma).

Let us take a closer look at **Greedy Step**. Consider an approximate two-player zero-sum matrix game: for every state $s \in \mathcal{S}$, we try to solve

$$\max_{x(\cdot|s) \in \Delta(\mathcal{A})} \min_{f(\cdot|s) \in \Delta(\mathcal{A})} x^\top A_s f, \tag{4}$$

$$A_s(a, b) = r(s, a, b) + \sum_{s'} \mathcal{P}(s' \mid s, a, b) V_{k-1}(s').$$

Here A_s represents a set of matrices related to current value function V_{k-1}. Instead of value-based approaches (PI (Patek, 1997), VI (Shapley, 1953)) which are often inefficient, we solve these matrix games by policy-based methods for efficiency and sub-optimality guarantee. We adopt the *Optimistic Mirror Descent* (Rakhlin and Sridharan, 2013) for two players by assuming access to population quantities, e.g., $A_s \pi$ in Eq. 4. Note that each $A_s(a, b) \in [0, \frac{1}{1-\gamma}]$ because $V_{k-1} \in [0, \frac{1}{1-\gamma})$.

For clarity, we follow notations in (Rakhlin and Sridharan, 2013). Denote $\phi(f, x) = x^\top A_s f$ which is convex w.r.t. f when fixing x and concave w.r.t. x when fixing f, and the domains for x, f are $\mathcal{X}, \mathcal{F}$ respectively. Thus $\mathcal{T} V_{k-1}(s) := \sup_{x \in \mathcal{X}} \inf_{f \in \mathcal{F}} \phi(f, x)$. we denote $\{y_t\}$ and $\{g_t\}$ as secondary sequences of $\{x_t\}$ and $\{f_t\}$ respectively. We refer readers to Appendix B for how we set the adaptive stepsizes η_t and η'_t.

We perform simultaneous updates for T' iterations in Algorithm 2 to minimize the following terms,

$$\frac{1}{T'} \sum_{t=1}^{T'} \phi(f_t, x_t) - \inf_f \sum_{t=1}^{T'} \frac{1}{T'} \phi(f, x_t), \tag{5}$$

$$\frac{1}{T'} \sum_{t=1}^{T'} (-\phi(f_t, x_t)) - \inf_x \frac{1}{T'} \sum_{t=1}^{T'} (-\phi(f_t, x)). \tag{6}$$

Suppose two infs are achieved at f^* and x^* respectively. With these two inequalities, we can derive an upper bound of **Greedy Step**:

$$\sup_{x \in \mathcal{X}} \inf_{f \in \mathcal{F}} \phi(f, x) - \inf_{f \in \mathcal{F}} \phi(f, x_{\bar{T}'})$$

to guarantee x^k is near-optimal with respect to V_{k-1}.

After obtaining x^k from **Greedy Step**, the **Iteration Step** aims to evaluate the value function while fixing $x = x^k$. We run T updates to find $f^* = \arg\min_f V^{x,f}$. In the competitive multi-agent RL literature, this step is equivalent to finding the *best response* of min player (namely, f^*) when fixing $x = x^k$. The intuition is that when the max player's policy is very close to its optimal policy at NE and f takes f^*, their accumulative value function is also close to V^* at NE. This step can be viewed as running NPG for a single-agent RL problem.

The following theorem gives the performance guarantee for Algorithm 1.

Theorem 1. *For Algorithm 1, set* $\eta \geq (1-\gamma)^2 \log |\mathcal{A}|$. *After* K *outer loops we have* $V^*(\rho) - \inf_f V^{x^K, f}(\rho)$ *upper bounded by*

$$\widetilde{O}\left(\frac{\mathcal{C}^{1,K,0}_{\rho,\sigma}}{(1-\gamma)^4 T} + \frac{\mathcal{C}^{0,K,0}_{\rho,\sigma}}{(1-\gamma)^4 T'} \log T' + \frac{\gamma^K}{1-\gamma} \mathcal{C}^{K,K+1,0}_{\rho,\sigma} \right).$$

We remind that σ is the optimization measure we use to train the policy and ρ is the performance measure of our interest.

Theorem 1 explicitly characterizes the performance of the output x^K in terms of the number of iterations and the concentrability coefficients. Viewing concentrability coefficients to be constants (which is the case when σ is sufficiently diverse) and looking at the dependency on T and K, we find the dependency on T is a fast $1/T$ rate, matching the same rate in the single agent NPG analysis (Agarwal et al., 2020). The dependency on K is exponential (γ^K) which means we only need a few outer loops. The first term has an $(1-\gamma)^{-4}$ dependency on the discount factor, which may not be tight and we leave it as a future work to improve. In Theorem 1 and 2, we set η to have a lower bound to simplify the convergence bounds. We can also derive η-dependent bounds. Note that the "large step size" phenomenon is also consistent with the single-agent setting (see Theorem 5.3 in (Agarwal et al., 2020) and discussion therein).

The proof of Theorem 1 further requires the following parts: mirror-descent type analysis of NPG used in (Agarwal et al., 2020) and simultaneous mirror descent for matrix games proposed in (Rakhlin and Sridharan, 2013). The full proof is deferred to Appendix B.

4.1 Extension: Entropy regularization

Following (Cen et al., 2020), we give an extension of entropy-regularized NPG in the **Iteration Step** for Algorithm 1. Denote τ as the regularization term, the entropy regularized value function is formulated as

$$V_\tau^{x,f}(\sigma) = V^{x,f}(\sigma) - \tau\mathcal{H}(\sigma, f) \tag{7}$$

where $\mathcal{H}(\sigma, f) = \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_\sigma^{x,f}}\mathbb{E}_{b\sim f}\log\frac{1}{f(b|s)}$ is the entropy term w.r.t. min player f. Note that $V_\tau^{x,f}(s) \in \left[-\tau\log|\mathcal{A}|, 1\right], \forall s \in \mathcal{S}$.

Entropy regularization requires us to minimize V_τ instead of original value function V. Denote $V_\tau^*(\sigma) = \min_f V_\tau^{x,f}(\sigma) = V^{x,f_\tau^*}(\sigma) - \tau\mathcal{H}(\sigma, f_\tau^*)$, the following sandwich bound holds

$$\begin{aligned} V^{x,f_\tau^*}(\sigma) &\geq V^{x,f^*(x)}(\sigma) \geq V_\tau^{x,f^*(x)}(\sigma) \\ &\geq V_\tau^*(\sigma) \geq V^{x,f_\tau^*}(\sigma) - \frac{\tau}{1-\gamma}\log|\mathcal{A}|. \end{aligned}$$

Therefore, the regularized problem and the original problem are close when τ is small.

NPG methods with entropy regularization. Let $\eta = \frac{1-\gamma}{\tau}$, we have the NPG update rule

$$\begin{aligned} &\theta^{t+1} \leftarrow \theta^t - \eta F_\sigma(\theta^t)^\dagger \nabla_\theta V_\tau^{x,f^t}(\sigma), \\ &f^{t+1}(b|s) \propto \exp\left(-\frac{1}{\tau}\sum_a x(a|s)Q_\tau^{x,f^t}(s,a,b)\right). \end{aligned}$$

The following theorem shows the performance improvement over Theorem 1.

Theorem 2. *For entropy regularized Algorithm 1, after K outer loops, one-sided measure $V^*(\rho) - \inf_f V^{x^K,f}(\rho)$ is bounded by*

$$\widetilde{O}\left(\frac{\gamma^T\mathcal{C}_{\rho,\sigma}^{1,K,0}}{(1-\gamma)^2}\left\|\frac{\sigma}{\mu_\tau^*}\right\|_\infty + \frac{\mathcal{C}_{\rho,\sigma}^{0,K,0}\log T'}{(1-\gamma)^4 T'} + \frac{\gamma^K\mathcal{C}_{\rho,\sigma}^{K,K+1,0}}{1-\gamma}\right).$$

Here, μ_τ^* is the one-sided *stationary distribution* w.r.t. x which satisfies: $\mu_\tau^* = d_{\mu_\tau^*}^{x,f_\tau^*}$. This argument indicates that the state visitation distribution remains unchanged if the initial state is already in a steady state. See Appendix B.1 for the full proof.

Recently, Perolat et al. (2021) studied learning algorithms for extensive-form zero-sum games and they also used entropy regularization. Compared with their work, the differences include 1) we use policy optimization instead of value-based methods used in their paper. 2) our entropy regularization is a simple extension whereas the entropy regularization is crucial in (Perolat et al., 2021): the regularization term gives strong convergence guarantees in monotone games.

5 Online Algorithm with Function Approximation

In this section, we extend Algorithm 1 to the realistic online setting with function approximation, in which the parameterization we adopt is defined in Section 3.2. In this setting, we only observe samples (instead of the population quantities in Section 4). The pseudo-code is listed in Algorithm 3.

To obtain estimates of quantities, we adopt the episodic sampling oracle (cf. Section 3) to provide transition tuples for estimating $A_s(a,b)$ in **Greedy Step** (cf. Eq. 4). This sampling oracle is also used in the **Iteration Step** to estimate value functions and gradients. See Appendix C for more details about how we use the sampling oracle.

In Section 2, we have pointed out that our algorithm has a smaller sample complexity of $O\left(\epsilon^{-6}\right)$ comparing to $O\left(\epsilon^{-12.5}\right)$ (Daskalakis et al., 2020). As for the computational complexity, we remark that we only need projecting onto an L_2-norm ball in Algorithm 3, 4, which has the same time complexity as computing the gradient (linear in the dimension of θ), so our algorithms are computationally efficient.

Now we describe our algorithm. Specifically, we let ξ and θ be parameters of x and f, respectively. [6] The

[6] We assume that the two players share the same pa-

output and motivation of the **Greedy Step** and the **Iteration Step** are analogous to those in Algorithm 1. In both steps, we need to take sample-based NPG updates which are forced to be constrained in a convex set $\mathcal{W} = \{w : \|w\|_2 \le W\}$ for analysis. From now, we denote W as the bound of this norm-constrained convex set where each NPG update lies.

Again, we first discuss the **Greedy Step** whose pseudo-code is listed in Algorithm 4. Our goal is still to obtain a near-optimal x^k with respect to $\mathcal{T}V_{k-1}$. Algorithm 4 is similar to Algorithm 2 in spirit. The main difference is that we use a sample-based NPG update rule for both x and f. Ideally, we wish to find simultaneous updates w_f^* and w_x^* for the players. Both are minimizers of quadratic loss

$$w_f^* = \arg\min_w \mathbb{E}_{s\sim\sigma}\mathbb{E}_{b\sim f^t(\cdot|s)} \left(w^\top \nabla_\theta \log f_\theta^t(b|s) - \left[(x_t^\top A_s)_b - \phi_s(f_t, x_t)\right]\right)^2.$$

$$w_x^* = \arg\min_w \mathbb{E}_{s\sim\sigma}\mathbb{E}_{a\sim x^t(\cdot|s)} \left(w^\top \nabla_\xi \log x_\xi^t(a|s) - \left[(A_s f_t)_a - \phi_s(f_t, x_t)\right]\right)^2.$$

Then the updates take the form $\theta_{t+1} = \theta_t - \eta w_f^*$ and $\xi_{t+1} = \xi_t + \eta w_x^*$. Along the way, the sampling oracle is used to approximate w_f^*, w_x^*. After T' iterations, we are able to output an approximate solution $x_{\bar{T}'}$ by averaging $\{x_t\}_{t=1}^{T'}$.

After obtaining x^k from the **Greedy Step**, we adapt NPG updates (Eq. 7) in Algorithm 1 to the online setting.

Denote $\nu^t = \nu_{\nu_0}^{x,f^t}$ for simplicity. Ideally, NPG update in the **Iteration Step** takes the form

$$w^t \in \arg\min_{\mathbb{E}_{s,a,b\sim\nu^t}} \left(w^\top \nabla_\theta \log f(b|s) - A^{x,f}(s,a,b)\right)^2,$$
$$\theta^{t+1} = \theta^t - \eta w^t. \tag{8}$$

We perform sample-based quadratic loss minimization, which shares similarity with the former step: it takes N steps of *projected gradient descent* to return an approximate update.

Now we state our main theorem.

Theorem 3. *For Algorithm 3, suppose in the **Greedy Step**: $\forall t \in [T'-1], \inf_{s,a} x^t(a|s), \inf_{s,b} f^t(b|s) \ge \iota^2$. Let $G = 4D(2DW + \frac{2}{1-\gamma})$. Set $\eta = \sqrt{\frac{2\log|\mathcal{A}|}{D^2W^2T}}, \eta' = \sqrt{\frac{2\log|\mathcal{A}|}{D^2W^2T'}}, \alpha = \frac{W}{G\sqrt{N}}, \alpha' = \frac{W}{G\sqrt{N'}}$. After K outer loops,*

rameter set only for clarity. We only need some minor modifications in the analysis to extend our results to the setting where two opposing players have different capabilities. Specifically, we only need to treat W (norm-bound of updates), D (regularity condition on features), and η separately for each agent.

Algorithm 3 Online Two-Player NPG

Input: $V_0 = 0$ value function
Output: Approximate policy x^K at NE
for $k = 1, 2, \cdots, K$ **do**
  Greedy Step:
  Run Algorithm 4 returns x^k with T' iterations.
  Iteration Step:
  Fix $x = x^k$, initialize $\theta^{(0)} = 0$.
  for $t = 0, 1, \cdots, T-1$ **do**
    Initialize $w_0 = 0$.
    for $n = 0, 1, \cdots, N-1$ **do**
      Sample $s, a, b \sim \nu^t$, then obtain $\hat{Q}(s,a,b)$ using the sampling oracle.
      Sample $b' \sim f^t(\cdot|s)$, observe:
      $g_n = \hat{Q}(s,a,b)(\nabla_\theta \log f^t(b|s) - \nabla_\theta \log f^t(b'|s))$.
      $w_{n+1} = \mathrm{Proj}_{\mathcal{W}}\big[w_n - 2\alpha\big(w_n^\top \nabla_\theta \log f^t(b|s) \nabla_\theta \log f^t(b|s) - g_n\big)\big]$.
    end for
    Set $\hat{w}^t = \frac{1}{N}\sum_{n=1}^N w_n$.
    Update $\theta^{(t+1)} = \theta^{(t)} - \eta \hat{w}^t$.
  end for
  Randomly sample f from $f^t (t = 0, 1 \cdots T-1)$.
  Denote V_k for $V^{x,f}$.
end for

$\mathbb{E}\left[V^*(\rho) - \inf_f V^{x^K,f}(\rho)\right]$ *is bounded by*

$$\widetilde{\mathcal{O}}\left(\frac{\mathcal{C}_{\rho,\sigma}^{1,K,0}}{(1-\gamma)^2}\epsilon + \frac{\mathcal{C}_{\rho,\sigma}^{0,K,0}}{(1-\gamma)^2}\epsilon' + \frac{\gamma^K}{1-\gamma}\mathcal{C}_{\rho,\sigma}^{K,K+1,0}\right)$$

where error terms ϵ, ϵ' are defined as

$$\epsilon = \sqrt{\frac{\log|\mathcal{A}|D^2W^2}{T}} + \frac{|\mathcal{A}|}{(1-\gamma)^2}\sqrt{\mathcal{C}'_{\sigma,\sigma}\frac{GW}{\sqrt{N}}} + \frac{|\mathcal{A}|}{(1-\gamma)^2}\sqrt{\mathcal{C}'_{\sigma,\sigma}\cdot\epsilon_{approx}}$$

$$\epsilon' = \sqrt{\frac{\log|\mathcal{A}|D^2W^2}{T'}} + \iota\left(\sqrt{\epsilon'_{approx}} + \frac{\sqrt{GW}}{N'^{\frac{1}{4}}}\right)$$

Here ϵ_{approx} and ϵ'_{approx} are *approximation errors* coming from **Greedy** and **Iteration Steps** (cf. Definition 2). We remind that D is a regularity condition on features, with which we could show log-linear parameterization is D^2-smooth (cf. Section 3.2). See Appendix C for specific expressions.

Similarly, the exponential γ^K in Theorem 3 implies that we only need a few outer iterations. When considering concentrability coefficients as constants, the dependency on T is a slower $T^{-1/2}$ rate while the sampling efficiency takes a $N^{-1/4}$ rate. Both match the rates in the sampling-based single-agent NPG analysis (Agarwal et al., 2020). We note that iteration counts T, T' and sample counts N, N' have the same exponent. There is no explicit dependence on

Algorithm 4 Online Greedy Step with Function-Approx

Input: $\theta_1, \xi_1 = \mathbf{0} \in \mathbb{R}^d$
Output: $x_{\bar{T'}}$ as average of $\{x_t\}, t \in [T']$
for $t = 1, 2, \cdots, T'$ **do**
min player: Initialize $w_0 = 0$.
for $n = 0, 1, 2 \cdots N' - 1$ **do**
Sample $s \sim \sigma(s), a \sim x^t(\cdot|s), b \sim f^t(\cdot|s), s' \sim \mathcal{P}(\cdot|s,a,b), b' \sim f^t(\cdot|s)$, observe:
$g_n = [r(s,a,b) + \gamma V_{k-1}(s')] \cdot (\nabla_\theta \log f^t(b|s) - \nabla_\theta \log f^t(b'|s))$.
Update: $w_{n+1} = \text{Proj}_{\mathcal{W}}[w_n - 2\alpha' \cdot (w_n^\top \nabla_\theta \log f^t(b|s) \nabla_\theta \log f^t(b|s) - g_n)]$.
end for
$\hat{w}^t = \frac{1}{N'} \sum_{n=1}^{N'} w_n$.
Update: $\theta_{t+1} = \theta_t - \eta' \hat{w}^t$.
max player: Initialize $w_0 = 0$.
for $n = 0, 1, 2 \cdots N' - 1$ **do**
Sample $s \sim \sigma(s), a \sim x^t(\cdot|s), b \sim f^t(\cdot|s), s' \sim \mathcal{P}(\cdot|s,a,b), a' \sim x^t(\cdot|s)$, observe:
$g_n = [r(s,a,b) + \gamma V_{k-1}(s')] \cdot (\nabla_\xi \log x^t(a|s) - \nabla_\xi \log x^t(a'|s))$.
Update: $w_{n+1} = \text{Proj}_{\mathcal{W}}\Big[w_n - 2\alpha' \cdot (w_n^\top \nabla_\xi \log x^t(a|s) \nabla_\xi \log x^t(a|s) - g_n)\Big]$.
end for
$\hat{w}^t = \frac{1}{N'} \sum_{n=1}^{N'} w_n$.
Update: $\xi_{t+1} = \xi_t + \eta' \hat{w}^t$.
end for

state-space $\mathcal{S}$ in the theorem, hence our online algorithm proves nice guarantees for function approximation even in the infinite-state setting. Instead, the bounds have parametric representation-related terms: D upper bounds feature norms $\|\phi_{s,a}\|$ and W restricts each NPG update. The term ι bounds two policy probabilities from below and it must be greater than 0 since we adopt log-linear parameterization. In spirit, ι is similar to concentrability coefficients which reflect the inherent dynamics of Markov games.

In the worst case, the concentrability coefficient scales as large as the number of states, and the bounds for function approximation are only meaningful in the benign case where the concentrability coefficient is small. However, we note that that, the dependency on concentrability is unavoidable: A hard example for the single-agent setting was given in (Agarwal et al., 2020). Since our Markov-Game (MG) setting is a generalization of the single-agent setting, their hard example also applies to our setting. Moreover, we argue that the coefficients can be small when there are some restrictions in the dynamics (see discussions in (Chen and Jiang, 2019)). We also use the same definition as in the prior work value-based learning (Perolat et al., 2015). The recent work (Daskalakis et al., 2020) also assumed this structure of MGs to analyze policy-based methods.

6 Conclusion

This paper gave the first quantitative analysis of policy gradient methods for general two-player zero-sum Markov games with function approximation. We quantified the performance gap of the output policy in terms of the number of iterations, number of samples, concentrability coefficients, and approximation error. An interesting direction is to extend our results to more advanced PG methods such as PPO (Schulman et al., 2017).

Acknowledgements

JDL acknowledges support of the ARO under MURI Award W911NF-11-1-0304, the Sloan Research Fellowship, NSF CCF 2002272, NSF IIS 2107304, and an ONR Young Investigator Award. SSD acknowledges funding from NSF Award's IIS-2110170 and DMS-2134106.

References

Alekh Agarwal, Sham M Kakade, Jason D Lee, and Gaurav Mahajan. Optimality and approximation with policy gradient methods in Markov decision processes. In *Conference on Learning Theory*, pages 64–66. PMLR, 2020.

András Antos, Csaba Szepesvári, and Rémi Munos. Learning near-optimal policies with Bellman-residual minimization based fitted policy iteration and a single sample path. *Machine Learning*, 71(1): 89–129, 2008.

Yakov Babichenko and Aviad Rubinstein. Communication complexity of approximate Nash equilibria. *Games and Economic Behavior*, 2020.

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

Yu Bai, Chi Jin, and Tiancheng Yu. Near-optimal reinforcement learning with self-play. *arXiv preprint arXiv:2006.12007*, 2020.

Marc Bellemare, Will Dabney, Robert Dadashi, Adrien Ali Taiga, Pablo Samuel Castro, Nicolas Le Roux, Dale Schuurmans, Tor Lattimore, and Clare Lyle. A geometric perspective on optimal representations for reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 4358–4369, 2019.

Jalaj Bhandari and Daniel Russo. Global optimal-

ity guarantees for policy gradient methods. *arXiv preprint arXiv:1906.01786*, 2019.

S. R. K. Branavan, Harr Chen, Luke S. Zettlemoyer, and Regina Barzilay. Reinforcement learning for mapping instructions to actions. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1 - Volume 1*, ACL '09, page 82–90, USA, 2009. Association for Computational Linguistics.

George W Brown. Iterative solution of games by fictitious play. *Activity analysis of production and allocation*, 13(1):374–376, 1951.

Noam Brown and Tuomas Sandholm. Superhuman AI for heads-up no-limit poker: Libratus beats top professionals. *Science*, 359(6374):418–424, 2018.

Noam Brown and Tuomas Sandholm. Solving imperfect-information games via discounted regret minimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 1829–1836, 2019.

Jingjing Bu, Lillian J Ratliff, and Mehran Mesbahi. Global convergence of policy gradient for sequential zero-sum linear quadratic dynamic games. *arXiv preprint arXiv:1911.04672*, 2019.

Shicong Cen, Chen Cheng, Yuxin Chen, Yuting Wei, and Yuejie Chi. Fast global convergence of natural policy gradient methods with entropy regularization. *arXiv preprint arXiv:2007.06558*, 2020.

Jinglin Chen and Nan Jiang. Information-theoretic considerations in batch reinforcement learning. In *International Conference on Machine Learning*, pages 1042–1051. PMLR, 2019.

Constantinos Daskalakis, Aranyak Mehta, and Christos Papadimitriou. Progress in approximate Nash equilibria. In *Proceedings of the 8th ACM conference on Electronic commerce*, pages 355–358, 2007.

Constantinos Daskalakis, Andrew Ilyas, Vasilis Syrgkanis, and Haoyang Zeng. Training GANs with optimism. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=SJJySbbAZ.

Constantinos Daskalakis, Dylan J Foster, and Noah Golowich. Independent policy gradient methods for competitive reinforcement learning. *arXiv preprint arXiv:2101.04233*, 2020.

Argyrios Deligkas, John Fearnley, Rahul Savani, and Paul Spirakis. Computing approximate Nash equilibria in polymatrix games. *Algorithmica*, 77(2):487–514, 2017.

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

Jakob N Foerster, Richard Y Chen, Maruan Al-Shedivat, Shimon Whiteson, Pieter Abbeel, and Igor Mordatch. Learning with opponent-learning awareness. *arXiv preprint arXiv:1709.04326*, 2017.

Kevin Gimpel and Noah A Smith. Softmax-margin crfs: Training log-linear models with cost functions. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 733–736, 2010.

Mika Göös and Aviad Rubinstein. Near-optimal communication lower bounds for approximate Nash equilibria. In *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 397–403. IEEE, 2018.

Xiaoxiao Guo, Satinder Singh, Richard Lewis, and Honglak Lee. Deep learning for reward design to improve monte carlo tree search in atari games. *arXiv preprint arXiv:1604.07095*, 2016.

Nicolas Heess, David Silver, and Yee Whye Teh. Actor-critic reinforcement learning with energy-based policies. In *European Workshop on Reinforcement Learning*, pages 45–58. PMLR, 2013.

Johannes Heinrich, Marc Lanctot, and David Silver. Fictitious self-play in extensive-form games. In *International Conference on Machine Learning*, pages 805–813, 2015.

Nan Jiang, Alex Kulesza, and Satinder Singh. Abstraction selection in model-based reinforcement learning. In *International Conference on Machine Learning*, pages 179–188, 2015.

Sham M Kakade. A natural policy gradient. In *Advances in neural information processing systems*, pages 1531–1538, 2002.

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

Edward Lockhart, Marc Lanctot, Julien Pérolat, Jean-Baptiste Lespiau, Dustin Morrill, Finbarr Timbers, and Karl Tuyls. Computing approximate equilibria in sequential adversarial games by exploitability descent. *arXiv preprint arXiv:1903.05614*, 2019.

Ishai Menache, Shie Mannor, and Nahum Shimkin. Basis function adaptation in temporal difference reinforcement learning. *Annals of Operations Research*, 134(1):215–238, 2005.

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*, 2019. URL https://openreview.net/forum?id=Bkg8jjC9KQ.

Seyed Sajad Mousavi, Michael Schukat, and Enda Howley. Traffic light control using deep policy-gradient and value-function-based reinforcement learning. *IET Intelligent Transport Systems*, 11(7): 417–423, 2017.

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

Stephen David Patek. *Stochastic and shortest path games: theory and algorithms*. PhD thesis, Massachusetts Institute of Technology, 1997.

Julien Perolat, Bruno Scherrer, Bilal Piot, and Olivier Pietquin. Approximate dynamic programming for two-player zero-sum Markov games. In *International Conference on Machine Learning*, pages 1321–1329, 2015.

Julien Pérolat, Bilal Piot, Bruno Scherrer, and Olivier Pietquin. On the use of non-stationary strategies for solving two-player zero-sum Markov games. In *AISTATS*, pages 893–901, 2016.

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*, pages 8525–8535. PMLR, 2021.

Sasha Rakhlin and Karthik Sridharan. Optimization, learning, and games with predictable sequences. In *Advances in Neural Information Processing Systems*, pages 3066–3074, 2013.

Julia Robinson. An iterative method of solving a game. *Annals of mathematics*, pages 296–301, 1951.

Bruno Scherrer. Approximate policy iteration schemes: a comparison. In *International Conference on Machine Learning*, pages 1314–1322, 2014.

Bruno Scherrer, Victor Gabillon, Mohammad Ghavamzadeh, and Matthieu Geist. Approximate modified policy iteration. *arXiv preprint arXiv:1205.3054*, 2012.

Bruno Scherrer, Mohammad Ghavamzadeh, Victor Gabillon, Boris Lesner, and Matthieu Geist. Approximate modified policy iteration and its application to the game of tetris. *J. Mach. Learn. Res.*, 16: 1629–1676, 2015.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International conference on machine learning*, pages 1889–1897, 2015.

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

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

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

Lloyd S Shapley. Stochastic games. *Proceedings of the national academy of sciences*, 39(10):1095–1100, 1953.

David Silver, Guy Lever, Nicolas Heess, Thomas Degris, Daan Wierstra, and Martin Riedmiller. Deterministic policy gradient algorithms. In *International conference on machine learning*, pages 387–395. PMLR, 2014.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of Go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

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.

Jiaming Song, Hongyu Ren, Dorsa Sadigh, and Stefano Ermon. Multi-agent generative adversarial imitation learning. In *Advances in neural information processing systems*, pages 7461–7472, 2018.

Richard S Sutton, David A McAllester, Satinder P Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in neural information processing systems*, pages 1057–1063, 2000.

Yuandong Tian, Jerry Ma, Qucheng Gong, Shubho Sengupta, Zhuoyuan Chen, James Pinkerton, and Larry Zitnick. Elf opengo: An analysis and open reimplementation of alphazero. In *International Conference on Machine Learning*, pages 6244–6253. PMLR, 2019.

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.

Henry G Weller, Gavin Tabor, Hrvoje Jasak, and Christer Fureby. A tensorial approach to computational continuum mechanics using object-oriented techniques. *Computers in physics*, 12(6):620–631, 1998.

Ming Yu, Zhuoran Yang, Mengdi Wang, and Zhaoran Wang. Provable q-iteration with l infinity guarantees and function approximation. In *Workshop on Optimization and RL, NeurIPS*, 2019.

Kaiqing Zhang, Zhuoran Yang, and Tamer Basar. Policy optimization provably converges to Nash equilibria in zero-sum linear quadratic games. In *Advances in Neural Information Processing Systems*, pages 11602–11614, 2019.

Kaiqing Zhang, Alec Koppel, Hao Zhu, and Tamer Basar. Global convergence of policy gradient methods to (almost) locally optimal policies. *SIAM Journal on Control and Optimization*, 58(6):3586–3612, 2020.

Martin Zinkevich, Michael Johanson, Michael Bowling, and Carmelo Piccione. Regret minimization in games with incomplete information. In *Advances in neural information processing systems*, pages 1729–1736, 2008.

A Basic results

In this section we provide some fundamental results for two-player zero-sum games and policy gradients.

Lemma 1 (contraction of Bellman operator)**.** *We show $\mathcal{T}_x v = \inf_f \mathcal{T}_{x,f} v$ is a γ contractor to V^x. Other forms of Bellman operators defined in Section 3 could be shown to hold contraction property with similar lines.*

Proof. First we show V^x is the unique fix point of $\mathcal{T}_x$, this is because:

$$\begin{aligned} V^x(s) &= r(s, x(s), f^*(s)) + \gamma \sum_{s'} \mathcal{P}(s'|s, x(s), f^*(s)) V^x(s') \\ &= \inf_f r(s, x(s), f(s)) + \gamma \sum_{s'} \mathcal{P}(s'|s, x(s), f(s)) V^x(s') \\ &= \mathcal{T}_x V^x(s) \end{aligned}$$

Then for all function $v : \mathbb{R}^{|\mathcal{S}|} \to \mathbb{R}^{|\mathcal{S}|}$,

$$\begin{aligned} & |\mathcal{T}_x v(s) - \mathcal{T}_x V^x(s)| \\ =& \left| \inf_f \mathcal{T}_{x,f} v(s) - \inf_f \mathcal{T}_{x,f} V^x(s) \right| \\ =& \max \left\{ \inf_f \mathcal{T}_{x,f} v(s) - \inf_f \mathcal{T}_{x,f} V^x(s), \inf_f \mathcal{T}_{x,f} V^x(s) - \inf_f \mathcal{T}_{x,f} v(s) \right\} \end{aligned}$$

Note that the first term could be upper bounded by

$$\begin{aligned} & \inf_f \mathcal{T}_{x,f} v(s) - \inf_f \mathcal{T}_{x,f} V^x(s) \\ =& \inf_f \mathcal{T}_{x,f} v(s) - \mathcal{T}_{x,f^*} V^x(s) \\ \le& \mathcal{T}_{x,f^*} v(s) - \mathcal{T}_{x,f^*} V^x(s) \\ =& \gamma \sum_{s'} \mathcal{P}(s'|s, x(s), f^*(s)) \left(v(s) - V^x(s) \right) \\ \le& \gamma \left| v(s) - V^x(s) \right| \end{aligned}$$

The second term could be upper bounded similarly, hence we have

$$|\mathcal{T}_x v - \mathcal{T}_x V^x| \le \gamma |v - V^x|$$

A direct application of contraction is $(\mathcal{T}_x)^\infty v = V^x$, which inspires the classical *Value Iteration* algorithm (Shapley, 1953). □

To analyze our NPG algorithm based on approximate dynamic programming scheme, we introduce the following lemma to upper bounding global performance, which is very useful in other sections.

Lemma 2. *Let ρ and σ be distributions over states. With Algorithm 1, after k iterations*

$$V^*(\rho) - \inf_f V^{x^k, f}(\rho) \le \frac{2(\gamma - \gamma^k) \mathcal{C}_{\rho,\sigma}^{1,k,0}}{(1-\gamma)^2} \epsilon + \frac{(1-\gamma^k) \mathcal{C}_{\rho,\sigma}^{0,k,0}}{(1-\gamma)^2} \epsilon' + \frac{2\gamma^k \mathcal{C}_{\rho,\sigma}^{k,k+1,0}}{1-\gamma}, \tag{9}$$

where

$$\begin{aligned} \epsilon &= \sup_{1 \le j \le k-1} \|\epsilon_j\|_{1,\sigma}, \\ \epsilon' &= \sup_{1 \le j \le k} \|\epsilon'_j\|_{1,\sigma}. \end{aligned}$$

This Lemma could be directly extended to expectation form in Section C.

This is a straightforward application of the following theorem.
(Perolat et al., 2015, Theorem 1) Let ρ and σ be distributions over states. Let p, q and q' be such that $\frac{1}{q}+\frac{1}{q'}=1$. Approximate Generalized Policy Iteration takes the following update:

$$\mathcal{T}V_{k-1} \leq \mathcal{T}_{x^k}V_{k-1}+\epsilon'_k \tag{10}$$
$$V_k = (\mathcal{T}_{x^k})^m V_{k-1}+\epsilon_k \tag{11}$$

Then, after k iterations, we have:

$$\begin{aligned}\|l_k\|_{p,\rho} \leq & \frac{2(\gamma-\gamma^k)(\mathcal{C}_q^{1,k,0})^{\frac{1}{p}}}{(1-\gamma)^2}\sup_{1\leq j\leq k-1}\|\epsilon_j\|_{pq',\sigma},\\ &+\frac{(1-\gamma^k)(\mathcal{C}_q^{0,k,0})^{\frac{1}{p}}}{(1-\gamma)^2}\sup_{1\leq j\leq k}\|\epsilon'_j\|_{pq',\sigma},\\ &+\frac{2\gamma^k}{1-\gamma}(\mathcal{C}_q^{k,k+1,0})^{\frac{1}{p}}\min(\|d_0\|_{pq',\sigma},\|b_0\|_{pq',\sigma}).\end{aligned}$$

where

$$\begin{aligned}\mathcal{C}_q^{l,k,d} &= \frac{(1-\gamma)^2}{\gamma^l-\gamma^k}\sum_{i=l}^{k-1}\sum_{j=i}^{\infty}c_q(j+d)\\ l_k &= V^*-\inf_f V^{x^k,f}\\ b_k &= V_k-\mathcal{T}_{x^{k+1}}V_k.\end{aligned}$$

Note the generalized norm of *Radon-Nikodym* derivative is:

$$c_q(j)=\sup_{\mu_1,\nu_1,\cdots,\mu_j,\nu_j}\left\|\frac{d(\rho\mathcal{P}_{\mu_1,\nu_1}\cdots\mathcal{P}_{\mu_j,\nu_j})}{d\sigma}\right\|_{q,\sigma}$$

Now we make adaptation to this theorem.

Proof. Set norm order $p=1$, then let $q\to\infty, q'=1$. Note that, in reinforcement learning, ρ has an explicit meaning of measure distribution or distribution for testing, while σ stands for exploration distribution. Normally exploration should cover more states, e.g., σ is a uniform distribution over all actions.
Now we provide detailed calculations with these parameter settings.

$$\begin{aligned}c_{q\to\infty}(j) &= \sup_{x^1,f^1,\cdots x^j,f^j}\left\|\frac{\rho\mathcal{P}_{x^1,f^1}\cdots\mathcal{P}_{x^j,f^j}}{\sigma}\right\|_{q\to\infty,\sigma}\\ &= \lim_{q\to\infty}\left(\sum_s\sigma(s)\left\|\frac{\rho\mathcal{P}_{x^1,f^1}\cdots\mathcal{P}_{x^j,f^j}(s)}{\sigma(s)}\right\|^q\right)^{\frac{1}{q}}\\ &= \sup_{x^1,f^1,\cdots x^j,f^j}\left\|\frac{\rho\mathcal{P}_{x^1,f^1}\cdots\mathcal{P}_{x^j,f^j}}{\sigma}\right\|_{\infty}\\ &= c_{\rho,\sigma}(j)\end{aligned}$$

As for weighted norm σ, it holds:

$$\begin{aligned}\|l_k\|_{1,\rho} &= \sum_s\rho(s)(V^*(s)-\inf_f V^{x^k,f}(s))\\ &= V^*(\rho)-\inf_f V^{x^k,f}(\rho)\end{aligned}$$

Notice in practice $V_0(s)$ is initialized to be 0, hence

$$\begin{aligned}\|b_0\|_{1,\sigma} &= \sum_s \sigma(s)|b_0(s)| \\ &= \sum_s \sigma(s)|V_0(s) - \mathcal{T}_{x^1} V_0(s)| \\ &\le \sup_{s,a,b} r(s,a,b) \\ &\le 1\end{aligned}$$

which gives that $\min(\|l_0\|_{1,\sigma}, \|b_0\|_{1,\sigma}) \le 1$. Then proof is completed via substitution. □

Lemma 3 (Policy Gradient)**.** *Consider a two-player zero-sum Markov game, when x is fixed, for f it holds:*

$$\begin{aligned}\nabla_\theta V^{x,f}(s_0) &= \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_{s_0}^{x,f}}\mathbb{E}_{a\sim x(\cdot|s)}\mathbb{E}_{b\sim f(\cdot|s)}\nabla_\theta \log f(b|s) Q^{x,f}(s,a,b) \\ &= \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_{s_0}^{x,f}}\mathbb{E}_{a\sim x(\cdot|s)}\mathbb{E}_{b\sim f(\cdot|s)}\nabla_\theta \log f(b|s) A^{x,f}(s,a,b)\end{aligned}$$

Proof. The proof is straightforward.

$$\begin{aligned}&\nabla_\theta V^{x,f}(s_0) \\ &= \nabla_\theta \left[\sum_{a_0} x(a_0|s_0) \sum_{b_0} f(b_0|s_0) Q^{x,f}(s_0,a_0,b_0)\right] \\ &= \sum_{b_0} \nabla_\theta f(b_0|s_0) \cdot \sum_{a_0} x(a_0|s_0) Q^{x,f}(s_0,a_0,b_0) + \mathbb{E}_{a_0}\mathbb{E}_{b_0}\nabla_\theta Q^{x,f}(s_0,a_0,b_0) \\ &= \mathbb{E}_{a_0}\mathbb{E}_{b_0}\left[\nabla_\theta \log f(b_0|s_0) Q^{x,f}(s_0,a_0,b_0)\right] + \gamma \mathbb{E}_{a_0}\mathbb{E}_{b_0}\mathbb{E}_{s_1}\nabla_\theta V^{x,f}(s_1) \\ &= \mathbb{E}_{x,f}\left[\sum_{t=0}^{\infty} \gamma^t \nabla_\theta \log f(b_t|s_t) Q^{x,f}(s_t,a_t,b_t)\right] \\ &= \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_{s_0}^{x,f}}\mathbb{E}_{a\sim x(\cdot|s)}\mathbb{E}_{b\sim f(\cdot|s)}\nabla_\theta \log f(b|s) Q^{x,f}(s,a,b)\end{aligned}$$

Notice that, when replacing $Q^{x,f}(s,a,b)$ in the final line with $A^{x,f}(s,a,b)$,

$$\begin{aligned}&\mathbb{E}_{a\sim x(\cdot|s)}\mathbb{E}_{b\sim f(\cdot|s)}\nabla_\theta \log f(b|s) A^{x,f}(s,a,b) \\ &= \mathbb{E}_{a\sim x(\cdot|s)}\mathbb{E}_{b\sim f(\cdot|s)}\nabla_\theta \log f(b|s)(Q^{x,f}(s,a,b) - V^{x,f}(s))\end{aligned}$$

Note that $\mathbb{E}_{b\sim f(\cdot|s)}\nabla_\theta \log f(b|s) = 0$, then $V^{x,f}(s)$ term's influence is zero. Proof is completed. □

The distribution mismatch coefficient, which is often used for single-agent policy-based optimzation, is a weaker condition compared to concentrability coefficients, see (Scherrer, 2014) for more discussion.

Lemma 4 (distribution mismatch coefficient and concentrability coefficients)**.** *For any fix policy x and its best response f^*, for infinite horizon, it holds*

$$\left\|\frac{d_\sigma^{x,f^*}}{\sigma}\right\|_\infty \le \frac{1}{1-\gamma}\mathcal{C}'_{\sigma,\sigma}$$

Proof. The proof is straightforward,

$$\begin{aligned}\left\|\frac{d_\sigma^{x,f^*}}{\sigma}\right\|_\infty &= (1-\gamma)\left\|\sum_{m\geq 0}\frac{\gamma^m\sigma(\mathcal{P}^{x,f^*})^m}{\sigma}\right\|_\infty \\ &\leq (1-\gamma)\sum_{m\geq 0}\gamma^m\cdot\left\|\frac{\gamma^m\sigma(\mathcal{P}^{x,f^*})^m}{\sigma}\right\|_\infty \\ &\leq (1-\gamma)\sum_{m\geq 1} m\gamma^{m-1}c_{\sigma,\sigma}(m-1) \\ &\leq \frac{\mathcal{C}'_{\sigma,\sigma}}{1-\gamma}\end{aligned}$$

Proof is completed. □

B Proof for Section 4

[Proof sketch] Recall Approximate Value/Polity Iteration for zero-sum games (Perolat et al., 2015), in each step $k = 1, 2, 3\cdots$,

$$\mathcal{T}V_{k-1} \leq \mathcal{T}_{x^k}V_{k-1} + \epsilon'_k \tag{12}$$
$$V_k = (\mathcal{T}_{x^k})^m V_{k-1} + \epsilon_k \tag{13}$$

where m denotes the number of performing Bellman operators w.r.t. a fixed value-function V_{k-1}, and ϵ_k is tolerance term. Specifically, $m = 1$ is Value Iteration. While $m \to \infty$, $(\mathcal{T}_{x^k})^m V_{k-1} \to V^{x^k} = \inf_f V^{x^k,f}$ due to Bellman operator's contraction property (see Lemma 1).

We discuss details to characterize error brought by two steps in each iteration, namely: ϵ'_k and ϵ_k respectively.

Greedy Step Suppose two sequences of $\{f_t\}$ and $\{x_t\}$ is given, let $\bar{f_{T'}} = \frac{1}{T'}\sum_{t=1}^{T'} f_t, \bar{x_T} = \frac{1}{T'}\sum_{t=1}^{T'} x_t$, then

$$\inf_f \frac{1}{T'}\sum_{t=1}^{T'}\phi(f,x_t) \leq \inf_f \phi(f,\bar{x_{T'}}) \leq \sup_x \inf_f \phi(f,x) \leq \sup_x \phi(\bar{f_{T'}},x) \leq \sup_x \frac{1}{T'}\sum_{t=1}^{T'}\phi(f_t,x), \tag{14}$$

where $\phi(f,x) = x^\top Af$ in this paper. Specifically, assume that two players produce sequences by using a regret minimization algorithm respectively, of which the bounds are

$$\frac{1}{T'}\sum_{t=1}^{T'}\phi(f_t,x_t) - \inf_f \sum_{t=1}^{T'}\frac{1}{T'}\phi(f,x_t) \leq Rate(x_1,x_2,\cdots x_{T'}) \tag{15}$$
$$\frac{1}{T'}\sum_{t=1}^{T'}(-\phi(f_t,x_t)) - \inf_x \frac{1}{T'}\sum_{t=1}^{T'}(-\phi(f_t,x)) \leq Rate(f_1,f_2,\cdots f_{T'}) \tag{16}$$

Thus, an upper bound of **Greedy Step** could be derived

$$\sup_x \inf_f \phi(f,x) - \inf_f \phi(f,\bar{x_{T'}}) \leq Rate(x_1,x_2,\cdots x_{T'}) + Rate(f_1,f_2,\cdots f_{T'})$$

Only in this section we denote (f^*,x^*) as the policy pair at NE for simplicity, i.e., $\sup_x \phi(f^*,x) = \inf_f \sup_x \phi(f,x)) = \inf_f \phi(f,x^*)$.

Lemma 5 (Greedy step suboptimality.)**.** *In Algorithm 2, when both players adopt the following adaptive step sizes for each state $s \in \mathcal{S}$,*

$$\eta_t^s = \min\left\{\frac{\log(|\mathcal{A}|T'^2)}{\sqrt{\sum_{i=1}^{t-1}\|A_s^\top x_i(\cdot|s) - A_s^\top x_{i-1}(\cdot|s)\|_\star^2} + \sqrt{\sum_{i=1}^{t-2}\|A_s^\top x_i(\cdot|s) - A_s^\top x_{i-1}(\cdot|s)\|_\star^2}}, \frac{1}{1+\frac{10}{(1-\gamma)^2}}\right\},$$
$$\eta_t^{s\prime} = \min\left\{\frac{\log(|\mathcal{A}|T'^2)}{\sqrt{\sum_{i=1}^{t-1}\|A_s f_i(\cdot|s) - A_s f_{i-1}(\cdot|s)\|_\star^2} + \sqrt{\sum_{i=1}^{t-2}\|A_s f_i(\cdot|s) - A_s f_{i-1}(\cdot|s)\|_\star^2}}, \frac{1}{1+\frac{10}{(1-\gamma)^2}}\right\},$$

then pair $(\bar{x_{T'}}, \bar{f_{T'}})$ *is an* $\tilde{O}\left(\frac{\log|\mathcal{A}|+\log T'}{(1-\gamma)^2 T'}\right)-$ *approximate minimax equilibrium.*

Proof. Proof is modified from (Rakhlin and Sridharan, 2013). Regret minimization procedure in Eq. 15 is calculated as:

$$\begin{aligned}
&\sum_{t=1}^{T'} \sup_x \phi(f_t, x) - \sup_x \phi(f^*, x) \\
\leq &\sum_{t=1}^{T'} \left\langle f_t - f^*, \nabla_f \left(\sup_x \phi(f_t, x) \right) \right\rangle \\
\leq &\left(\frac{1}{\eta_1}\right) R_{max}^2 + \sum_{t=1}^{T'} \|A^\top x_t - A^\top x_{t-1}\|_\star \|g_t - f_t\| \\
&- \frac{1}{2}\sum_{t=1}^{T'} \frac{1}{\eta_t}(\|g'_t - f_t\|^2 + \|g'_{t-1} - f_t\|) + 1,
\end{aligned}$$

where R_{max}^2 is upper bound of KL divergence between f^* and any g', so $R_{max}^2 = \log(|\mathcal{A}|T'^2)$. With some calculations, the sum of two regrets (Eq. 15 and its counterpart) is upper bounded by

$$6 + \left(4 + \frac{40}{(1-\gamma)^2}\right)\log(|\mathcal{A}|T'^2) + \frac{1}{T'}\frac{40}{(1-\gamma)^2} \tag{17}$$

Thus regret is upper bounded by $\mathcal{O}\left(\frac{\log|\mathcal{A}|+\log T'}{(1-\gamma)^2 T'}\right)$.

□

Iteration Step While Approximate Value-based algorithms generally focus on the relation between ϵ_k and accumulative error of $\epsilon_{k,i}, i = 1, 2, 3 \cdots m$. We could take another view of this iteration: at k^{th} iteration, let V^{x^k} be our goal to achieve, V_k is what we finally get with optimization techniques, and ϵ_k now turns out to be a suboptimality gap.

$$\sum_s |\epsilon_k(s)|\sigma(s) = V^{x, f^T}(\sigma) - \inf_f V^{x^k, f}(\sigma)$$

Describe this with general policy-based languages: when we are given x^k, we desire to find $\inf_f V^{x^k,f}$. Note that it holds similarity to the general single-agent MDP, where the agent seeks to find $\max_\pi V^\pi$. Thus we could apply NPG for player two at iteration step, next we show NPG update in Algorithm 1 takes exponential form on the weighted advantage function.

Lemma 6. *Fix* x*, when* f *is softmax parameterized, it holds*

$$f^{t+1} \propto f^t \cdot \exp^{-\frac{\eta}{1-\gamma}\sum_a x(a|s)A^{x,f}(s,a,b)}$$

Proof. Notice that Fisher matrix calculation for this case is often obtained by minimizing

$$L(w) = \mathbb{E}_{s\sim d_\sigma^{x,f}} \mathbb{E}_{b\sim f}\left(w^\top \nabla_\theta \log f(b|s) - \sum_a x(a|s) A^{x,f}(s,a,b)\right)^2,$$

which is because:
At minimizer w^*, $\frac{dL(w^*)}{dw} = 0$ implies that

$$\mathbb{E}_{s\sim d_\sigma^{x,f}} \mathbb{E}_{b\sim f}\left((w^*)^\top \nabla_\theta \log f(b|s) - \sum_a x(a|s) A^{x,f}(s,a,b)\right) \nabla_\theta \log f(b|s) = 0,$$

rearrange this,

$$w^* = (1-\gamma) F_\sigma(\theta)^\dagger \nabla_\theta V(\sigma)$$

For softmax parameterization, note:

$$w^* = \sum_a x(a|s)A^{x,f}(s,a,b) + v(s) \Leftrightarrow L(w^*) = 0$$

Then NPG updates take the following form,

$$\theta^{t+1} = \theta^t - \frac{\eta}{1-\gamma}\sum_a x(a|s)A^{x,f}(s,a,b) - \frac{\eta}{1-\gamma}v$$
$$f^{t+1} \propto f^t \cdot \exp^{-\frac{\eta}{1-\gamma}\sum_a x(a|s)A^{x,f}(s,a,b)}$$

Proof is completed. □

Lemma 7. *With above update rule, after T iterations*

$$V^{x,f^T}(\sigma) - V^{x,f^*}(\sigma) \leq \frac{\log|\mathcal{A}|}{\eta T} + \frac{1}{(1-\gamma)^2 T},$$

where f^* is player two's best response w.r.t. x, which satisfies $V^{x,f^*}(\sigma) = \inf_f V^{x,f}(\sigma)$

Proof. Proof sketch is analogous to Agarwal et al. (2020, Theorem 5.3), only need to replace $A^\pi(s,a)$ with an average term $\sum_a x(a|s)A^{x,f}(s,a,b)$. □

With these policy optimization results, the proof of Theorem 1 is concise.

[Proof for Theorem 1]

Proof. After T steps of NPG descent, it can be guaranteed that

$$\epsilon_k(s) = V^{x^k,f^T}(s) - \inf_f V^{x^k,f} > 0$$
$$\sum_s |\epsilon_k(s)|\sigma(s) = V^{x,f^T}(\sigma) - \inf_f V^{x^k,f}(\sigma) \leq \frac{2}{(1-\gamma)^2 T},$$

where we have set $\eta \geq (1-\gamma)^2 \log|\mathcal{A}|$. Substitute this NPG suboptimality and Greedy step suboptimality (Lemma 5) into Lemma 2, proof is completed. □

B.1 Proof for Entropy regularization

First, note that the entropy term w.r.t. min player f

$$\mathcal{H}(\sigma, f) = \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_\sigma^{x,f}}\mathbb{E}_{b\sim f}\log\frac{1}{f(b|s)}$$

lies in $\left[0, \log|\mathcal{A}|\right]$. Recall that $V_\tau^*(\sigma) = \min_f V_\tau^{x,f}(\sigma) = V^{x,f_\tau^*}(\sigma) - \tau\mathcal{H}(\sigma, f_\tau^*)$ and the following sandwich bound holds:

$$V^{x,f_\tau^*}(\sigma) \geq V^{x,f^*(x)}(\sigma) \geq V_\tau^{x,f^*(x)}(\sigma) \geq V_\tau^*(\sigma) \geq V^{x,f_\tau^*}(\sigma) - \frac{\tau}{1-\gamma}\log|\mathcal{A}|,$$

We aim to bound $V^{x,f^T}(\sigma) - V^{x,f^*(x)}$ through optimizing V_τ, denote $V_\tau^* = V^{x,f^*(x)}$ for short, observe

$$\begin{aligned} V^{x,f^T}(\sigma) - V^{x,f^*(x)}(\sigma) &= V^{x,f^T}(\sigma) - V_\tau^{x,f^T}(\sigma) + V_\tau^{x,f^T}(\sigma) - V_\tau^*(\sigma) \\ &\leq \frac{\tau}{1-\gamma}\log|\mathcal{A}| + V_\tau^{x,f^T}(\sigma) - V_\tau^*(\sigma) + 0. \end{aligned}$$

Besides the notations in Section 4, introduce the regularized advantage function for min player f

$$A_\tau^{x,f}(s,a,b) = Q_\tau^{x,f}(s,a,b) + \tau \log f(b|s) - V_\tau^{x,f}(s)$$

Regularized reward is

$$r_\tau(s,a,b) = r(s,a,b) + \tau \log f(b|s)$$

Lemma 8 (Regularized policy gradients)**.**

$$\nabla_\theta V_\tau^{x,f}(s_0) = \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_{s_0}^{x,f}}\mathbb{E}_{a\sim x}\mathbb{E}_{b\sim f}\nabla_\theta \log f(b|s) A_\tau^{x,f}(s,a,b)$$

Proof. Note that soft Q function is $Q_\tau^{x,f}(a,a,b) = r(s,a,b) + \gamma \mathbb{E}_{s'} V_\tau^{x,f}(s')$

$$\begin{aligned}
\nabla_\theta V_\tau^{x,f}(s_0) &= \nabla_\theta \left[\sum_{b_0} f(b_0|s_0) \sum_{a_0} x(a_0|s_0)\Big(r(s_0,a_0,b_0) + \tau \log f(b_0|s_0) + \gamma \mathbb{E}_{s_1} V_\tau^{x,f}(s_1)\Big)\right] \\
&= \nabla_\theta \left[\sum_{b_0} f(b_0|s_0)\mathbb{E}_{a_0\sim x}\Big(Q_\tau^{x,f}(s_0,a_0,b_0) + \tau \log f(b_0|s_0)\Big)\right] \\
&= \sum_{b_0} f(b_0|s_0)\nabla_\theta \log f(b_0|s_0)\mathbb{E}_{a_0\sim x}\Big(Q_\tau^{x,f}(s_0,a_0,b_0) + \tau \log f(b_0|s_0)\Big) \\
&\qquad + \mathbb{E}_{b_0\sim f}\mathbb{E}_{a_0\sim x}\nabla_\theta\Big(r(s_0,a_0,b_0) + \gamma \mathbb{E}_{s_1} V_\tau^{x,f}(s_1) + \tau \log f(b_0|s_0)\Big) \\
&= \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \nabla_\theta \log f(b_t|s_t)\Big(Q_\tau^{x,f}(s_t,a_t,b_t) + \tau \log f(b_t|s_t)\Big)\right] \\
&= \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_{s_0}^{x,f}}\mathbb{E}_{a\sim x}\mathbb{E}_{b\sim f}\nabla_\theta \log f(b|s)\Big(Q_\tau^{x,f}(s,a,b) + \tau \log f(b|s)\Big) \\
&= \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_{s_0}^{x,f}}\mathbb{E}_{a\sim x}\mathbb{E}_{b\sim f}\nabla_\theta \log f(b|s) A_\tau^{x,f}(s,a,b)
\end{aligned}$$

□

Adopting softmax parameterization, gradient is written as

$$\frac{\partial V_\tau^{x,f}(s_0)}{\partial \theta(s,b)} = \frac{1}{1-\gamma} d_{s_0}^{x,f}(s) f(b|s)\mathbb{E}_{a\sim x} A_\tau^{x,f}(s,a,b)$$

Lemma 9 (Regularized update rule)**.**

$$f^{t+1}(b|s) \propto \big(f^t(b|s)\big)^{1-\frac{\eta\tau}{1-\gamma}} \exp\Big(-\frac{\eta}{1-\gamma}\sum_a x(a|s) Q_\tau^{x,f}(s,a,b)\Big) \tag{18}$$

Proof. Denote update direction as $w^* = (F_\sigma^\theta)^\dagger \nabla_\theta V_\tau^{x,f}(\sigma)$, which means w^* is minimizer of square loss

$$\begin{aligned}
&\left\|F_\sigma^\theta w - \nabla_\theta V_\tau^{x,f}(\sigma)\right\|^2 \\
=& \sum_{s,b}\left(d_\sigma^{x,f}(s) f(b|s)(w_{s,b} - c(s)) - \frac{1}{1-\gamma} d_\sigma^{x,f} f(b|s)\mathbb{E}_{a\sim x} A_\tau^{x,f}(s,a,b)\right)^2,
\end{aligned}$$

thus $w_{s,b} = c(s) + \frac{1}{1-\gamma}\mathbb{E}_{a\sim x} A^{x,f}(s,a,b)$, and

$$\begin{aligned}
f^{t+1}(b|s) &\propto f^t(b|s)\exp\left(-\frac{\eta}{1-\gamma}\mathbb{E}_{a\sim x} A_\tau^{(t)}(s,a,b)\right) \\
&= f^t(b|s)\exp\left(-\frac{\eta}{1-\gamma}\mathbb{E}_{a\sim x}(Q_\tau^{(t)}(s,a,b) + \tau \log f(b|s) - V_\tau^{(t)}(s))\right) \\
&\propto \big(f^t(b|s)\big)^{1-\frac{\eta\tau}{1-\gamma}}\exp\left(-\frac{\eta}{1-\gamma}\sum_a x(a|s) Q_\tau^{(t)}(s,a,b)\right)
\end{aligned}$$

Proof is completed. □

Lemma 10 (Performance improvement lemma)**.**

$$V_\tau^t(s_0) = V_\tau^{t+1}(s_0) + \mathbb{E}_{s\sim d_{s_0}^{t+1}}\left[(\frac{1}{\eta} - \frac{\tau}{1-\gamma})KL\left(f^{t+1}(\cdot|s)\|f^t(\cdot|s)\right) + \frac{1}{\eta}KL\left(f^t(\cdot|s)\|f^{t+1}(\cdot|s)\right)\right]$$

Proof. Regularized update rule can be transformed to

$$\frac{1-\gamma}{\eta}\left(\log f^{t+1}(b|s) - \log f^t(b|s)\right) + \frac{1-\gamma}{\eta}\log Z^t(s) = -\tau\log f^t(b|s) - \mathbb{E}_{a\sim x}Q_\tau^t(s,a,b)$$

Then

$$\begin{aligned}
V_\tau^t(s_0) &= \mathbb{E}_{a\sim x}\mathbb{E}_{b\sim f^t}\left[\tau\log f^t(b_0|s_0) + Q_\tau^t(s_0,a_0,b_0)\right] \\
&= -\frac{1-\gamma}{\eta}\log Z^t(s_0) + \frac{1-\gamma}{\eta}KL\left(f^t(\cdot\|s_0), f^{t+1}(\cdot\|s_0)\right) \\
&= \mathbb{E}_{b_0\sim f^{t+1}}\left[\tau\log f^t(b|s) + \mathbb{E}_{a\sim x}Q_\tau^t(s,a,b) + \frac{1-\gamma}{\eta}\left(\log f^{t+1}(b|s) - \log f^t(b|s)\right)\right] \\
&\quad + \frac{1-\gamma}{\eta}KL\left(f^t(\cdot|s_0)\|f^{t+1}(\cdot|s_0)\right) \\
&= \mathbb{E}_{b_0\sim f^{t+1}}\left[\tau\log f^{t+1}(b_0|s_0) + \mathbb{E}_{a\sim x}Q_\tau^t(s_0,a_0,b_0)\right] \\
&\quad + \left(\frac{1-\gamma}{\eta} - \tau\right)KL\left(f^{t+1}(\cdot|s_0)\|f^t(\cdot|s_0)\right) + \frac{1-\gamma}{\eta}KL\left(f^t(\cdot|s_0)\|f^{t+1}(\cdot|s_0)\right)
\end{aligned}$$

Note that: $Q_\tau^t(s_0,a_0,b_0) = r(s_0,a_0,b_0) + \gamma\mathbb{E}_{s_1}V_\tau^t(s_1)$, apply this recurrently then the proof is completed. □

For regularized Markov games, the suboptimality gap is shown to be

$$\begin{aligned}
&V_\tau^{x,f_\tau^*}(\sigma) - V_\tau^{x,f^t}(\sigma) \\
&= \mathbb{E}\left[\sum_{i=0}^\infty \gamma^i\left(r(s_i,a_i,b_i) + \tau\log f_\tau^*(b_i|s_i)\right)\right] - V_\tau^{x,f^t}(\sigma) \\
&= \mathbb{E}\left[\sum_{i=0}^\infty \gamma^i\left(r(s_i,a_i,b_i) + \tau\log f_\tau^*(b_i|s_i) + \gamma V_\tau^{x,f^t}(s_{i+1}) - V_\tau^{x,f^t}(s_i)\right)\right] \\
&= \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_\sigma^{x,f_\tau^*}}\left[\sum_b f_\tau^*(b|s)\left(\mathbb{E}_{a\sim x}Q_\tau^{(t)}(s,a,b) + \tau\log f_\tau^*(b|s)\right) - V_\tau^{(t)}(s)\right].
\end{aligned}$$

Take reverse,

$$\begin{aligned}
&V_\tau^{x,f^t}(\sigma) - V_\tau^{x,f_\tau^*}(\sigma) \\
&= \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_\sigma^{x,f_\tau^*}}\left[V_\tau^{(t)}(s) + \sum_b f_\tau^*(b|s)\left(-\mathbb{E}_{a\sim x}Q_\tau^{(t)} - \tau\log f_\tau^*(b|s)\right)\right],
\end{aligned}$$

where

$$\begin{aligned}
&\sum_b f_\tau^*(b|s)\left(-\mathbb{E}_{a\sim x}Q_\tau^{(t)}(s,a,b) - \tau\log f_\tau^*(b|s)\right) \\
&= \tau\sum_b f_\tau^*(b|s)\log\left(\frac{e^{\mathbb{E}_{a\sim x}-Q_\tau^{(t)}(s,a,b)/\tau}}{f_\tau^*(b|s)}\right) \\
&\le \tau\log\sum_b\exp\left(-\mathbb{E}_{a\sim x}\frac{Q_\tau^{(t)}(s,a,b)}{\tau}\right)
\end{aligned}$$

Contraction property Following contraction argument raised by Cen et al. (2020), suppose $\eta = \frac{1-\gamma}{\tau}$

$$\begin{aligned}
&V_\tau^{x,f^{t+1}}(\sigma) - V_\tau^{x,f_\tau^*}(\sigma) \\
&= V_\tau^{x,f^{t+1}}(\sigma) - V_\tau^{x,f^t}(\sigma) + V_\tau^{x,f^t}(\sigma) - V_\tau^{x,f_\tau^*}(\sigma) \\
&= \mathbb{E}_{s\sim d_\rho^{t+1}} - \frac{1}{\eta} KL\left(f^t(\cdot|s) \| f^{t+1}(\cdot|s)\right) + V_\tau^{x,f^t}(\sigma) - V_\tau^{x,f_\tau^*}(\sigma) \\
&\le \left(V_\tau^{x,f^t}(\sigma) - V_\tau^{x,f_\tau^*}(\sigma)\right) \cdot \left(1 - \left\| \frac{d_\rho^{x,f_\tau^*}}{d_\rho^{t+1}} \right\|_\infty^{-1}\right) \\
&\le \left(V_\tau^{x,f^t}(\sigma) - V_\tau^{x,f_\tau^*}(\sigma)\right) \cdot \left[1 - (1-\gamma)\left\| \frac{d_\rho^{x,f_\tau^*}}{\rho} \right\|_\infty^{-1}\right]
\end{aligned}$$

Denote *stationary distribution* as: $\mu_\tau^* = d_{\mu_\tau^*}^{x,f_\tau^*}$ and

$$\begin{aligned}
&V_\tau^{x,f^t}(\sigma) - V_\tau^{x,f_\tau^*}(\sigma) \\
&\le \left\| \frac{\sigma}{\mu_\tau^*} \right\|_\infty \cdot \left(V_\tau^{x,f^t}(\mu_\tau^*) - V_\tau^{x,f_\tau^*}(\mu_\tau^*)\right) \\
&\le \left\| \frac{\sigma}{\mu_\tau^*} \right\|_\infty \cdot \gamma^t \left(V_\tau^{x,f^0}(\mu_\tau^*) - V_\tau^{x,f_\tau^*}(\mu_\tau^*)\right)
\end{aligned}$$

Combining these results, we are ready to show Theorem 2.

[Proof for Theorem 2]

Proof.

$$\begin{aligned}
&V^{x,f^T}(\sigma) - V^{x,f^*(x)}(\sigma) \\
&= V^{x,f^T}(\sigma) - V_\tau^{x,f^T}(\sigma) + V_\tau^{x,f^T}(\sigma) - V_\tau^*(\sigma) + V_\tau^*(\sigma) - V^{x,f^*(x)}(\sigma) \\
&\le \frac{\tau \log|\mathcal{A}|}{1-\gamma} + \left\| \frac{\sigma}{\mu_\tau^*} \right\|_\infty \cdot \gamma^T \left(V_\tau^{x,f^0}(\mu_\tau^*) - V_\tau^{x,f_\tau^*}(\mu_\tau^*)\right),
\end{aligned}$$

note that $V_\tau^{x,f^0}(\mu_\tau^*) - V_\tau^{x,f_\tau^*}(\mu_\tau^*) \le 1 + \tau \log|\mathcal{A}|$.

Proof is completed via substitution into Lemma 2. □

C Proof for Section 5

For online setting, we consider Approximate Generalized Policy Iteration (Eq.10, Eq.11) in expectation

$$\mathbb{E}[\mathcal{T} V_{k-1}] \le \mathbb{E}[\mathcal{T}_{x^k} V_{k-1}] + \mathbb{E}[\epsilon_k'] \tag{19}$$

$$\mathbb{E}[V_k] = \mathbb{E}\left[(\mathcal{T}_{x^k})^m V_{k-1}\right] + \mathbb{E}[\epsilon_k] \tag{20}$$

Here, we try to bound the summation of tolerances over state space $\mathcal{S}$. In function approximation, $\mathcal{S}$ could be very large or even infinite. Therefore, we use the optimization measure σ we take to train our policy for generalization across states, namely:

$$\mathbb{E}[\epsilon_k'] = \mathbb{E}[\sum_s \sigma(s)\epsilon_k'(s)]$$

$$\mathbb{E}[\epsilon_k] = \mathbb{E}[\sum_s \sigma(s)\epsilon_k(s)]$$

Randomness is brought by oracle sampling and stochastic optimization.

Based on the iterative scheme, Lemma 2 could be adapted to expectation: after k iterations,

$$\mathbb{E}\left[V^*(\rho) - \inf_f V^{x^k,f}(\rho)\right] \leq \frac{2(\gamma-\gamma^k)\mathcal{C}^{1,k,0}_{\rho,\sigma}}{(1-\gamma)^2}\epsilon + \frac{(1-\gamma^k)\mathcal{C}^{0,k,0}_{\rho,\sigma}}{(1-\gamma)^2}\epsilon' + \frac{2\gamma^k\mathcal{C}^{k,k+1,0}_{\rho,\sigma}}{1-\gamma},$$

where

$$\epsilon = \sup_{1\leq j\leq k-1}\mathbb{E}[\epsilon_j],$$
$$\epsilon' = \sup_{1\leq j\leq k}\mathbb{E}[\epsilon'_j].$$

We are able to derive suboptimality gap for the online setting.

[Proof sketch] Similar to Section B, discuss errors brought by two phases respectively.

Greedy Step

Problem Restatement: Consider a two-player zero-sum matrix game, formally

$$\min_{f(\cdot|s)\in\Delta(|\mathcal{A}|)}\max_{x(\cdot|s)\in\Delta(|\mathcal{A}|)} f^\top A_s x$$
$$A_s(a,b) = r(s,a,b) + \sum_{s'}\mathcal{P}(s' \mid s,a,b)V_{k-1}(s')$$

The goal is to output policy $x^-_{T'}$ for max player and an upper bound of **Greedy error** $\mathbb{E}[\epsilon'] = \mathbb{E}[\sum_s \sigma(s)\epsilon'_k(s)]$:

$$\mathbb{E}\left[\sum_s \sup_x \inf_f \phi_s(f(\cdot|s),x(\cdot|s)) - \inf_f \phi_s(f(\cdot|s),x^-_{T'}(\cdot|s))\right]$$
$$\leq \frac{1}{T'}\mathbb{E}\sum_s\left\{\sum_{t=1}^{T'}\phi_s(f_t,x_t) - \inf_f\sum_{t=1}^{T'}\phi_s(f,x_t) + \sum_{t=1}^{T'}(-\phi_s(f_t,x_t)) - \inf_x\sum_{t=1}^{T'}(-\phi_s(f_t,x))\right\}$$

We only analyze max player (x^t) and min player (f^t) is very similar.

First, we show our Algorithm 4 is using unbiased gradient estimates. Observe:

$$\begin{aligned}
&\mathbb{E}[g_n]\\
&= \mathbb{E}_{s\sim\sigma}\mathbb{E}_{a\sim x^t(\cdot|s)}\mathbb{E}_{b\sim f^t(\cdot|s)}\mathbb{E}_{s'\sim\mathcal{P}(\cdot|s,a,b)}\mathbb{E}_{a'\sim x^t(\cdot|s)}[r(s,a,b)+\gamma V_{k-1}(s')]\\
&\qquad\cdot(\nabla_\xi \log x^t(a|s) - \nabla_\xi \log x_t(a'|s))\\
&= \mathbb{E}_{s\sim\sigma}\mathbb{E}_{a\sim x^t(\cdot|s)}(A_s f^t)_a\nabla_\xi \log x^t(a|s) - \mathbb{E}_{s\sim\sigma}\mathbb{E}_{a'\sim x^t(\cdot|s)}\phi_s(f_t,x_t)\nabla_\xi\log x^t(a'|s)\\
&= \mathbb{E}_{s\sim\sigma}\mathbb{E}_{a\sim x^t(\cdot|s)}[(A_s f_t)_a - \phi_s(f_t,x_t)]\nabla_\xi\log x^t(a|s).
\end{aligned}$$

Recall notations raised in Eq. 5, where x^* is *best response* of average policy $\frac{1}{T'}\sum_{t=1}^{T'} f^t$,

$$\begin{aligned}
&\mathbb{E}_{s\sim\sigma}KL(x^*(\cdot|s)\|x^t(\cdot|s)) - KL(x^*(\cdot|s)\|x^{t+1}(\cdot|s))\\
&= \mathbb{E}_{s\sim\sigma}\mathbb{E}_{a\sim x^*}\log\frac{x^{t+1}(a|s)}{x^t(a|s)}\\
&\geq \mathbb{E}_{s\sim\sigma}\mathbb{E}_{a\sim x^*(\cdot|s)}\langle\nabla_\theta\log x^t(a|s),\eta'\hat{w}^t\rangle - \frac{\beta}{2}\|\xi^{t+1}-\xi^t\|^2\\
&= \eta'\mathbb{E}_{s\sim\sigma}\mathbb{E}_{a\sim x^*}\left[\nabla_\theta\log x^t(a|s)^\top\hat{w}^t - ((A_sf_t)_a - \phi_s(f_t,x_t))\right]\\
&\qquad+\eta'\mathbb{E}_{s\sim\sigma}\mathbb{E}_{a\sim x^*(\cdot|s)}\left[(A_sf_t)_a-\phi_s(f_t,x_t)\right] - \frac{\beta\eta'^2W^2}{2}\\
&\geq -\eta'\sqrt{\mathbb{E}_{s\sim\sigma}\mathbb{E}_{a\sim x^*(\cdot|s)}\left(\nabla_\theta\log x^t(a|s)^\top\hat{w}^t - [(A_sf_t)_a-\phi_s(f_t,x_t)]\right)^2}\\
&\qquad+\eta'\mathbb{E}_{s\sim\sigma}(\phi_s(f_t,x^*)-\phi_s(f_t,x_t)) - \frac{\beta\eta'^2W^2}{2},
\end{aligned}$$

where we define $L(w^t) = \mathbb{E}_{s\sim\sigma}\mathbb{E}_{a\sim x^t}\left(\nabla_\theta \log x^t(a|s)^\top w^t - [(A_s f_t)_a - \phi_s(f_t, x_t)]\right)^2$ with little abuse of notation.

Rearrange inequality and the upper bound of $\mathbb{E}_{s\sim\sigma}\phi_s(f_t, x^*) - \phi_s(f_t, x_t)$ is smaller than

$$\frac{1}{\eta'}\mathbb{E}_{s\sim\sigma}\left[KL(x^*(\cdot|s)\|x^t(\cdot|s)) - KL(x^*(\cdot|s)\|x^{t+1}(\cdot|s))\right] + \sqrt{\sup_{t\le T'}\left\|\frac{1}{x^t}\right\|_\infty}\sqrt{L(\hat{w}^t)} + \frac{\beta}{2}\eta' W^2$$

Expectation of ϵ_{est} is bounded by sample complexity, SGD optimizer has

$$\epsilon_{est} = \mathbb{E}[L(\hat{w}^t)] - L(w^*) \le \frac{GW}{\sqrt{N'}},$$

where $G = 2B(BW + 2/1-\gamma)$ bounds norm of gradient estimation, learning rate α' is set as $W/G\sqrt{N'}$, see Lemma 13.

Thus

$$\begin{aligned}
&\mathbb{E}\left[\sup_x \inf_f \phi_s(f(\cdot|s), x(\cdot|s)) - \inf_f \phi_s(f(\cdot|s), \bar{x_{T'}}(\cdot|s))\right] \\
\le\ & \frac{1}{\eta'}\frac{1}{T'}\mathbb{E}_{s\sim\sigma}\left[KL(x^*(\cdot|s)\|x^1(\cdot|s)) + KL(f^*(\cdot|s)\|f^1(\cdot|s))\right] \\
&+ \left(\sqrt{\sup_{t\le T'}\left\|\frac{1}{f^t}\right\|_\infty} + \sqrt{\sup_{t\le T'}\left\|\frac{1}{x^t}\right\|_\infty}\right)\cdot\sqrt{\epsilon'_{approx} + \frac{GW}{\sqrt{N'}}} + \beta\eta' W^2 \\
\le\ & \frac{2\log|\mathcal{A}|}{\eta' T'} + \left(\sqrt{\sup_{t\le T'}\left\|\frac{1}{f^t}\right\|_\infty} + \sqrt{\sup_{t\le T'}\left\|\frac{1}{x^t}\right\|_\infty}\right)\cdot\sqrt{\epsilon'_{approx} + \frac{GW}{\sqrt{N'}}} + \beta\eta' W^2
\end{aligned}$$

Let $\eta' = \sqrt{\frac{2\log|\mathcal{A}|}{\beta W^2 T'}}$, and finally $\mathbb{E}[\epsilon'] = \mathbb{E}[\sum_s \sigma(s)\epsilon'_k(s)]$ is lower than

$$2\sqrt{\frac{2\log|\mathcal{A}|\beta W^2}{T'}} + 2\iota\left(\sqrt{\epsilon'_{approx}} + \frac{\sqrt{GW}}{N'^{\frac{1}{4}}}\right) \tag{21}$$

Iteration Step See NPG regret Lemma 11 for two-player zero-sum games, where err_t is bounded when $\nu_0(s,a,b) = \sigma(s)/|\mathcal{A}|^2$ is an exploration distribution covering all states and actions:

$$\begin{aligned}
|err_t| &\le \sqrt{\mathbb{E}_{s\sim d_\sigma^{x,f^*}, a\sim x, b\sim f^*(x)}\left[A^{x,f^t}(s,a,b) - w^t\nabla_\theta \log f^t(b|s)\right]^2} \\
&\le \sqrt{\left\|\frac{d_\sigma^{x,f^*}\cdot x\cdot f^*}{\nu^t}\right\|_\infty \mathbb{E}_{s,a,b\sim\nu^t}\left(A^{x,f^t}(s,a,b) - w^t\nabla_\theta \log f^t(b|s)\right)^2} \\
&\le \sqrt{\frac{|\mathcal{A}|^2}{1-\gamma}\left\|\frac{d_\sigma^{x,f^*}}{\sigma}\right\|_\infty L(\hat{w}^t, \theta)},
\end{aligned}$$

Notice $\mathbb{E}\sqrt{\frac{1}{T}\sum_t L(\hat{w}^t,\theta)} \le \sqrt{\frac{1}{T}\sum_t \mathbb{E}[L(\hat{w}^t,\theta)]}$, then proof is completed via upper bounding $\mathbb{E}[L(\hat{w}^t,\theta)]$.

The final equality contains *distribution mismatch coefficient* $\left\|d_\sigma^{x,f^*}/\sigma\right\|_\infty$, which often appears in single-agent policy-based optimization. It measures the difficulty of exploration problems faced by algorithms. Furthermore, *concentrability coefficients* are stronger, from which $\left\|d_\sigma^{x,f^*}/\sigma\right\|_\infty$ could be derived. See Lemma 4.

We first introduce a two-player zero-sum Markov game version regret lemma, single agent version of MDP is useful for online NPG analysis (Agarwal et al., 2020).

Lemma 11 (NPG regret). *Assume for all $s \in \mathcal{S}$ and $b \in \mathcal{A}$ that $\log f(b|s)$ is a β-smooth function, then*

$$\frac{1}{T}\sum_{t=0}^{T-1} V^{x,f^t}(\sigma) - V^{x,f^*(x)}(\sigma) \le \frac{1}{1-\gamma}\left(\frac{\log|\mathcal{A}|}{\eta T} + \frac{\eta\beta W^2}{2} - \frac{1}{T}\sum_{t=0}^{T-1} err_t\right),$$

where err_t is defined as

$$\begin{aligned} err_t &= \mathbb{E}_{s\sim d_\sigma^{x,f^*(x)}} \mathbb{E}_{b\sim f^*(x)} \left[\sum_a x(a|s) A^{x,f^t}(s,a,b) - w^t \nabla_\theta \log f^t(b|s) \right] \\ &= \mathbb{E}^*_{s,a,b} \left[A^{x,f^t}(s,a,b) - w^t \nabla_\theta \log f^t(b|s) \right] \end{aligned}$$

where we denote $\mathbb{E}^*_{s,a,b} \coloneqq \mathbb{E}_{s\sim d_\sigma^{x,f^*}} \mathbb{E}_{a\sim x} \mathbb{E}_{b\sim f^*}$ for simplicity.

Proof. When making no abuse of notation, we denote f^* as the *best response* of fixed x for simplicity from now on, i.e., $V^{x,f^*} = \inf_f V^{x,f}$

$$\begin{aligned} &\mathbb{E}^*_{s,a,b} \left(KL(f^* \| f^t) - KL(f^* \| f^{t+1}) \right) \\ &= \mathbb{E}^*_{s,a,b} \log \frac{f^{t+1}(b|s)}{f^t(b|s)} \\ &\geq \mathbb{E}^*_{s,a,b} \left[-\eta \nabla_\theta \log f^t(b|s) w^t - \frac{\beta\eta^2}{2} W^2 \right] \\ &= -\eta \mathbb{E}^*_{s,a,b} A^{x,f^t}(s,a,b) + \eta \mathbb{E}^*_{s,a,b} \left(A^{x,f^*}(s,a,b) - \nabla_\theta \log f^t(b|s) \right) - \frac{\beta\eta^2 W^2}{2} \\ &= -\eta(1-\gamma) \left(V^{x,f^*}(\sigma) - V^{x,f^t}(\sigma) \right) + \eta \, err_t - \frac{\beta\eta^2 W^2}{2}. \end{aligned}$$

Rearrange it and we get

$$\begin{aligned} &V^{x,f^t}(\sigma) - V^{x,f^*}(\sigma) \\ &\leq \frac{1}{1-\gamma} \left(\frac{1}{\eta} \mathbb{E}_{s\sim d_\sigma^{x,f^*}} \mathbb{E}_{a\sim x} \left(KL(f^* \| f^t) - KL(f^* \| f^{t+1}) \right) - err_t + \frac{\eta\beta W^2}{2} \right) \end{aligned}$$

Taking the sum, and notice that $\theta^0 = 0$

$$\frac{1}{T} \sum_{t=0}^{T-1} (V^{x,f^t}(\sigma) - V^{x,f^*}(\sigma)) \leq \frac{1}{1-\gamma} \left(\frac{\log|\mathcal{A}|}{\eta T} + \frac{\eta\beta W^2}{2} - \frac{1}{T} \sum_{t=0}^{T-1} err_t \right)$$

□

Lemma 12 (Unbiased estimation)**.** *Sample-based gradient in Algorithm 3 is unbiased of $\nabla_w L(w)$ (Eq. 8).*

Proof. Recall the estimators in (Agarwal et al., 2020, Algorithm 1, 3) which provide unbiased estimations of $Q^{x,f^t}(s,a,b)$ and d_ν^{x,f^t}. With little abuse of notation, we use Q^t, A^t to represent Q^{x,f^t} and A^{x,f^t}.

$$\begin{aligned} &\mathbb{E}_{s,a,b\sim\nu^t} \mathbb{E}_{v'\sim f^t} [g_n] \\ &= \mathbb{E}_{s,a,b\sim\nu^t} \hat{Q}(s,a,b) \nabla_\theta \log f^t(b|s) - \mathbb{E}_{s,a,b\sim\nu^t} \mathbb{E}_{v'\sim f^t} \hat{Q}(s,a,b) \nabla_\theta \log f^t(v'|s) \\ &= \mathbb{E}_{s,a,b\sim\nu^t} Q^t(s,a,b) \nabla_\theta \log f^t(b|s) - \mathbb{E}_{s,a,b\sim\nu^t} V^t(s) \nabla_\theta \log f^t(b|s) \\ &= \mathbb{E}_{s,a,b\sim\nu^t} A^t(s,a,b) \nabla_\theta \log f^t(b|s), \end{aligned}$$

hence,

$$\begin{aligned} &2\mathbb{E}_{s,a,b\sim\nu^t} \left[\left(w_n^\top \nabla_\theta \log f^t(b|s) \right) \nabla_\theta \log f^t(b|s) - g_n \right] \\ &= 2\mathbb{E}_{s,a,b\sim\nu^t} \left[w_n^\top \nabla_\theta \log f^t(b|s) - A^t(s,a,b) \right] \nabla_\theta \log f^t(b|s) \\ &= \nabla_w L(w_n) \end{aligned}$$

Proof is completed.

□

Lemma 13 (Bounded stat error). *Assume $\|\nabla_\theta \log f(b|s)\|_2 \le B$, statistical error of minimizing Eq. 8 is bounded*

$$\mathbb{E}\left[L(\hat{w}^t)\right] - L(w^*) = \mathcal{O}(\frac{1}{\sqrt{N}})$$

Proof. For this sample-based projected gradient descent, notice the estimated gradient is bounded by $G \coloneqq 2B(BW + \frac{1}{1-\gamma})$. Shalev-Shwartz and Ben-David (2014) shows if setting learning rate $\alpha = \frac{W}{G\sqrt{N}}$,

$$\mathbb{E}[L(\bar{w})] - L(w^*) \le \frac{GW}{\sqrt{N}}$$

□

Lemma 14 (Gradient norm bounded for log-linear parameterization). *Suppose $\pi_\theta(a|s) = \frac{\exp(\theta^\top \phi_{s,a})}{\sum_{a' \in \mathcal{A}} \exp(\theta^\top \phi_{s,a'})}$ for which $\|\phi_{s,a}\| \le D$, we show $\|\nabla_\theta \log \pi(a|s)\|_2 \le B = 2D$.*

Proof. Proof is straight forward

$$\begin{aligned}
\nabla_\theta \log \pi_\theta(a|s) &= \phi_{s,a} - \frac{\phi_{s,a'} e^{\theta^\top \phi_{s,a}}}{\sum_{a'} e^{\theta^\top \phi_{s,a'}}} \\
&= \phi_{s,a} - \sum_{a'} \phi_{s,a'} P(a'),
\end{aligned}$$

where $P(a') = \frac{e^{\theta^\top \phi_{s,a}}}{\sum_{a'} e^{\theta^\top \phi_{s,a'}}}$. Then

$$\begin{aligned}
\|\nabla_\theta \log \pi_\theta(a|s)\| &\le \|\phi_{s,a}\| + \|\sum_{a'} \phi_{s,a'} P(a')\| \\
&\le \|\phi_{s,a}\| + \sum_{a'} P(a') \|\phi_{s,a'}\| \\
&\le 2D.
\end{aligned}$$

Proof is completed. □

Lemma 15 (Iteration error of Algorithm 3). *Set learning rate $\eta = \sqrt{\frac{2\log|\mathcal{A}|}{\beta T W^2}}$, $\alpha = \frac{W}{G\sqrt{N}}$, initial state-action distribution $\nu_0(s,a,b) = {}^{\sigma(s)}/_{|\mathcal{A}|^2}$, err_t in Lemma 11 can be bounded with sample complexity.*

$$\begin{aligned}
|err_t|^2 &\le \mathbb{E}^*_{s,a,b}\left[A^{x,f^t}(s,a,b) - w^t \nabla_\theta \log f^t(b|s)\right]^2 \\
&\le \left\| \frac{d_\sigma^{x,f^*} \cdot x \cdot f^*}{\nu^t} \right\|_\infty \mathbb{E}_{s,a,b\sim\nu^t}\left(A^{x,f^t}(s,a,b) - w^t \nabla_\theta \log f^t(b|s)\right)^2 \\
&\le \frac{1}{1-\gamma} \left\| \frac{d_\sigma^{x,f^*} \cdot x \cdot f^*}{\nu_0} \right\|_\infty L(\hat{w}^t, \theta) \\
&\le \frac{1}{1-\gamma} \left\| \frac{d_\sigma^{x,f^*} \cdot x \cdot f^*}{\nu_0} \right\|_\infty L(\hat{w}^t, \theta) \\
&\le \frac{|\mathcal{A}|^2}{1-\gamma} \left\| \frac{d_\sigma^{x,f^*}}{\sigma} \right\|_\infty L(\hat{w}^t, \theta) \\
&\le \frac{|\mathcal{A}|^2}{1-\gamma} \left\| \frac{d_\sigma^{x,f^*}}{\sigma} \right\|_\infty \left(L(\hat{w}^t) - L(w^*) + L(w^*)\right)
\end{aligned}$$

From Lemma 4, $\left\| \frac{d_\sigma^{x,f^*}}{\sigma} \right\|_\infty$ is controlled by $\mathcal{C}'_{\sigma,\sigma}$

Take the expectation on both sides of Lemma 11, summation of err_t is bounded

$$\begin{aligned}
\mathbb{E}\left[\sum_t \frac{-1}{T} err_t\right] &\leq \mathbb{E}\left[\frac{1}{T}\sum_t \sqrt{\mathbb{E}^*_{s,a,b}\left(A^{x,f^t}(s,a,b) - w^t \nabla_\theta \log f^t(b|s)\right)^2}\right] \\
&\leq \mathbb{E}\sqrt{\frac{1}{T}\sum_t \mathbb{E}^*_{s,a,b}\left(A^{x,f^t}(s,a,b) - w^t \nabla_\theta \log f^t(b|s)\right)^2},\ y=\sqrt{x}\ \textit{is concave} \\
&\leq \sqrt{\frac{1}{T}\sum_t \mathbb{E}\left[\mathbb{E}^*_{s,a,b}\left(A^{x,f^t}(s,a,b) - w^t \nabla_\theta \log f^t(b|s)\right)^2\right]} \\
&\leq \sqrt{\frac{|\mathcal{A}|^2}{1-\gamma}\left\|\frac{d_\sigma^{x,f^*}}{\sigma}\right\|_\infty \cdot \mathbb{E}\left[L(\hat{w}^t) - L(w^*) + L(w^*)\right]} \\
&\leq \sqrt{\frac{|\mathcal{A}|^2}{(1-\gamma)^2}\mathcal{C}'_{\sigma,\sigma}\left(\frac{GW}{\sqrt{N}} + \epsilon_{approx}\right)}
\end{aligned}$$

Further, $\forall 1 \leq j \leq k-1$, it holds

$$\begin{aligned}
&\mathbb{E}[\sum_s \sigma(s)\epsilon_j(s)] \\
&= \mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}(V^{x,f^t}(\sigma) - V^{x,f^*}(\sigma))\right] \\
&\leq \sqrt{\frac{2\log|\mathcal{A}|\beta W^2}{T}} + \frac{|\mathcal{A}|}{(1-\gamma)^2}\sqrt{\mathcal{C}'_{\sigma,\sigma}\left(\frac{GW}{\sqrt{N}} + \epsilon_{approx}\right)} \\
&\leq \sqrt{\frac{2\log|\mathcal{A}|\beta W^2}{T}} + \frac{|\mathcal{A}|}{(1-\gamma)^2}\sqrt{\mathcal{C}'_{\sigma,\sigma}\frac{GW}{\sqrt{N}}} + \frac{|\mathcal{A}|}{(1-\gamma)^2}\sqrt{\mathcal{C}'_{\sigma,\sigma}\cdot\epsilon_{approx}}
\end{aligned}$$

Take the expectation on both sides of Lemma 2, note $\mathbb{E}[\sup_{1\leq j\leq k-1}\|\epsilon_j\|_{1,\sigma}]$ is also upper bounded by the above inequality, then the proof is completed via substitution.

Combining these results, Theorem 3 for online setting is concluded.

[Proof for Theorem 3]

Proof. Substitute ϵ and ϵ',

$$\begin{aligned}
&\mathbb{E}\left[V^*(\rho) - \inf_f V^{x^k,f}(\rho)\right] \\
\leq &\frac{2(\gamma-\gamma^k)\mathcal{C}^{1,k,0}_{\rho,\sigma}}{(1-\gamma)^2}\cdot\epsilon + \frac{(1-\gamma^k)\mathcal{C}^{0,k,0}_{\rho,\sigma}}{(1-\gamma)^2}\cdot\epsilon' + \frac{2\gamma^k}{1-\gamma}\mathcal{C}^{,k+1,0}_{\rho,\sigma},
\end{aligned}$$

where

$$\begin{aligned}
\epsilon &= \sqrt{\frac{2\log|\mathcal{A}|\beta W^2}{T}} + \frac{|\mathcal{A}|}{(1-\gamma)^2}\sqrt{\mathcal{C}'_{\sigma,\sigma}\frac{GW}{\sqrt{N}}} + \frac{|\mathcal{A}|}{(1-\gamma)^2}\sqrt{\mathcal{C}'_{\sigma,\sigma}\cdot\epsilon_{approx}} \\
\epsilon' &= 2\sqrt{\frac{2\log|\mathcal{A}|\beta W^2}{T'}} + 2\iota\left(\sqrt{\epsilon'_{approx}} + \frac{\sqrt{GW}}{N'^{\frac{1}{4}}}\right)
\end{aligned}$$

When the outer loop count k is set as K, proof of Theorem 3 is completed. ∎