Supplementary Materials: "Provable Self-Play Algorithms for Competitive Reinforcement Learning"

Yu Bai [*1] **Chi Jin** [*2]

A. Proofs for Section 3

A.1. Proof of Theorem 2

Notation: To be clear from the context, we denote the upper bound and lower bound Q^{up} and Q^{low} computed at the k-th episode as $Q^{\mathsf{up},k}$ and $Q^{\mathsf{low},k}$, and policies computed and used at the k-th episode as μ^k and ν^k.

Choice of bonus: $\beta_t = c\sqrt{SH^2\iota/t}$ for sufficient large absolute constant c.

Lemma 1 (ULCB). *With probability at least $1-p$, we have following bounds for any (s,a,b,h,k):*

$$V_h^{\mathsf{up},k}(s) \geq \sup_\mu V_h^{\mu,\nu^k}(s), \qquad Q_h^{\mathsf{up},k}(s,a,b) \geq \sup_\mu Q_h^{\mu,\nu^k}(s,a,b) \tag{1}$$

$$V_h^{\mathsf{low},k}(s) \leq \inf_\nu V_h^{\mu^k,\nu}(s), \qquad Q_h^{\mathsf{low},k}(s,a,b) \leq \inf_\nu Q_h^{\mu^k,\nu}(s,a,b) \tag{2}$$

Proof. By symmetry, we only need to prove the statement (1). For each fixed k, we prove this by induction from $h=H+1$ to $h=1$. For base case, we know at the $(H+1)$-th step, $V_{H+1}^{\mathsf{up},k}(s) = \sup_\mu V_{H+1}^{\mu,\nu^k}(s) = 0$.

Now, assume the left inequality in (1) holds for $(h+1)$-th step, for the h-th step, we first recall the updates for Q functions respectively:

$$\begin{aligned} Q_h^{\mathsf{up},k}(s,a,b) =& \min\left\{r_h(s,a,b) + [\widehat{\mathbb{P}}_h^k V_{h+1}^{\mathsf{up},k}](s,a,b) + \beta_t, H\right\} \\ \sup_\mu Q_h^{\mu,\nu^k}(s,a,b) =& r_h(s,a,b) + [\mathbb{P}_h \sup_\mu V_{h+1}^{\mu,\nu^k}](s,a,b) \end{aligned}$$

In case of $Q_h^{\mathsf{up},k}(s,a,b) = H$, the right inequality in (1) clearly holds. Otherwise, we have:

$$\begin{aligned} Q_h^{\mathsf{up},k}(s,a,b) - \sup_\mu Q_h^{\mu,\nu^k}(s,a,b) =& [\widehat{\mathbb{P}}_h^k V_{h+1}^{\mathsf{up},k}](s,a,b) - [\mathbb{P}_h^k \sup_\mu V_{h+1}^{\mu,\nu^k}](s,a,b) + \beta_t \\ =& [\widehat{\mathbb{P}}_h^k (V_{h+1}^{\mathsf{up},k} - \sup_\mu V_{h+1}^{\mu,\nu^k})](s,a,b) - [(\widehat{\mathbb{P}}_h^k - \mathbb{P}_h)\sup_\mu V_{h+1}^{\mu,\nu^k}](s,a,b) + \beta_t \end{aligned}$$

Since $\widehat{\mathbb{P}}_h^k$ preserves the positivity, by induction assumption, we know the first term is positive. By Lemma 2, we know the second term $\geq -\beta_t$. This finishes the proof of the right inequality in (1).

To prove the left inequality in (1), again recall the updates for V functions respectively:

$$\begin{aligned} V_h^{\mathsf{up},k}(s) =& \mu_h^k(s)^\top Q_h^{\mathsf{up},k}(s,\cdot,\cdot)\nu_h^k(s) = \max_{\phi\in\Delta_{\mathcal{A}}} \phi^\top Q_h^{\mathsf{up},k}(s,\cdot,\cdot)\nu_h^k(s) \\ \sup_\mu V_h^{\mu,\nu^k}(s) =& \max_{\phi\in\Delta_{\mathcal{A}}} \phi^\top [\sup_\mu Q_h^{\mu,\nu^k}(s,\cdot,\cdot)]\nu_h^k(s) \end{aligned}$$

[*]Equal contribution [1]Salesforce Research [2]Princeton University. Correspondence to: Yu Bai <yu.bai@salesforce.com>, Chi Jin <chij@princeton.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

where the first equation is by the definition of policy μ^k the algorithm picks. Therefore:

$$V_h^{\mathsf{up},k}(s) - \sup_\mu V_h^{\mu,\nu^k}(s) \geq \max_{\phi\in\Delta_{\mathcal{A}}} \phi^\top [Q_h^{\mathsf{up},k} - \sup_\mu Q_h^{\mu,\nu^k}](s,\cdot,\cdot)\nu_h^k(s) \geq 0.$$

This finishes the proof. □

Lemma 2 (Uniform Concentration)**.** *Consider value function class*

$$\mathcal{V}_{h+1} = \{V : \mathcal{S}_{h+1} \to \mathbb{R} \mid V(s) \in [0,H] \text{ for all } s \in \mathcal{S}_{h+1}\}.$$

There exists an absolute constant c, with probability at least $1-p$, we have:

$$\left|[(\hat{\mathbb{P}}_h^k - \mathbb{P}_h)V](s,a,b)\right| \leq c\sqrt{SH^2\iota/N_h^k(s,a,b)} \quad \text{for all } (s,a,b,k,h) \text{ and all } V \in \mathcal{V}_{h+1}.$$

Proof. We show this for one (s,a,b,k,h); the rest follows from a union bound over these indices (and results in a larger logarithmic factor.) Throughout this proof we let $c>0$ to be an absolute constant that may vary from line to line.

Let $\mathcal{V}_\varepsilon$ be an ε-covering of $\mathcal{V}_{h+1}$ in the ∞ norm (that is, for any $V \in \mathcal{V}_{h+1}$ there exists $\widehat{V} \in \mathcal{V}_\varepsilon$ such that $\sup_s |V(s) - \widehat{V}(s)| \leq \varepsilon$.) We have $|\mathcal{V}_\varepsilon| \leq (1/\varepsilon)^S$, and by Hoeffding inequality and a union bound (over both $\widehat{V}$ and $N_h^k \in [K]$), we have with probability at least $1-p$ that

$$\left|\sup_{\widehat{V}\in\mathcal{V}_\varepsilon}\left[(\hat{\mathbb{P}}_h^k - \mathbb{P}_h)\widehat{V}\right]\right| \leq \sqrt{\frac{H^2(S\log(1/\varepsilon)+\log(K/p))}{N_h^k(s,a,b)}}.$$

Taking $\varepsilon = c\sqrt{H^2S\iota/K}$, the above implies that

$$\left|\sup_{\widehat{V}\in\mathcal{V}_\varepsilon}\left[(\hat{\mathbb{P}}_h^k - \mathbb{P}_h)\widehat{V}\right]\right| \leq c\sqrt{\frac{H^2S\iota}{N_h^k(s,a,b)}}.$$

Meanwhile, with this choice of ε, for any $V \in \mathcal{V}_{h+1}$, there exists $\widehat{V} \in \mathcal{V}_\varepsilon$ such that $\sup_s |V(s) - \widehat{V}(s)| \leq \varepsilon$, and therefore

$$\left|\left[(\hat{\mathbb{P}}_h^k - \mathbb{P}_h)V\right] - \left[(\hat{\mathbb{P}}_h^k - \mathbb{P}_h)\widehat{V}\right]\right| \leq 2\varepsilon = c\sqrt{\frac{H^2S\iota}{K}} \leq c\sqrt{\frac{H^2S\iota}{N_h^k(s,a,b)}}.$$

Combining the preceding two bounds, we have that the desired concentration holds for all $V \in \mathcal{V}_{h+1}$.

□

Proof of Theorem 2. By Lemma 1, we know the regret,

$$\text{Regret}(K) = \sum_{k=1}^K \left[\sup_\mu V_1^{\mu,\nu^k}(s_1^k) - \inf_\nu V_1^{\mu^k,\nu}(s_1^k)\right] \leq \sum_{k=1}^K [V_1^{\mathsf{up},k}(s_1^k) - V_1^{\mathsf{low},k}(s_1^k)]$$

On the other hand, by the updates in Algorithm 1, we have:

$$\begin{aligned}
[V_h^{\mathsf{up},k} - V_h^{\mathsf{low},k}](s_h^k) =& \mu_h^k(s_h^k)^\top [Q_h^{\mathsf{up},k} - Q_h^{\mathsf{low},k}](s_h^k,\cdot,\cdot)\nu_h^k(s_h^k),\\
=& [Q_h^{\mathsf{up},k} - Q_h^{\mathsf{low},k}](s_h^k,a_h^k,b_h^k) + \xi_h^k\\
\leq& [\widehat{\mathbb{P}}_h^k(V_{h+1}^{\mathsf{up},k} - V_{h+1}^{\mathsf{low},k})](s_h^k,a_h^k,b_h^k) + 2\beta_h^k + \xi_h^k\\
\leq& [\mathbb{P}(V_{h+1}^{\mathsf{up},k} - V_{h+1}^{\mathsf{low},k})](s_h^k,a_h^k,b_h^k) + 4\beta_h^k + \xi_h^k\\
=& (V_{h+1}^{\mathsf{up},k} - V_{h+1}^{\mathsf{low},k})(s_{h+1}^k) + 4\beta_h^k + \xi_h^k + \zeta_h^k
\end{aligned}$$

the last inequality is due to Lemma 2. (Recall that $\beta_h^k := \beta_{N_h^k(s_h^k,a_h^k,b_h^k)} = c\sqrt{H^2S\iota/N_h^k(s_h^k,a_h^k,b_h^k)}$ when $N_h^k \geq 1$. In the case when $N_h^k = 0$, we can still define $\beta_h^k = \beta_0 := c\sqrt{H^2S\iota}$, and the above inequality still holds as we have $Q_h^{\mathsf{up},k} - Q_h^{\mathsf{low},k} = H \leq \beta_0$.) Above, ξ_h^k and ζ_h^k are defined as

$$\begin{aligned}\xi_h^k =& \mathbb{E}_{a\sim\mu_h^k(s_h^k),b\sim\nu_h^k(s_h^k)}[Q_h^{\mathsf{up},k} - Q_h^{\mathsf{low},k}](s_h^k,a,b) - [Q_h^{\mathsf{up},k} - Q_h^{\mathsf{low},k}](s_h^k,a_h^k,b_h^k)\\ \zeta_h^k =& \mathbb{E}_{s\sim\mathbb{P}_h(\cdot|s_h^k,a_h^k,b_h^k)}[(V_{h+1}^{\mathsf{up},k} - V_{h+1}^{\mathsf{low},k})](s) - [V_{h+1}^{\mathsf{up},k} - V_{h+1}^{\mathsf{low},k})](s_{h+1}^k)\end{aligned}$$

Both ξ_h^k and ζ_h^k are martingale difference sequence, therefore by the Azuma-Hoeffding inequality we have with probability $1-p$ that

$$\sum_{k,h}\xi_h^k \leq \mathcal{O}(\sqrt{HT\iota}) \quad \text{and} \quad \sum_{k,h}\zeta_h^k \leq \mathcal{O}(\sqrt{HT\iota}).$$

Therefore, by our choice of bonus β_t and the Pigeonhole principle, we have

$$\begin{aligned}&\sum_{k=1}^K\left[V_1^{\mathsf{up},k}(s_1^k) - V_1^{\mathsf{low},k}(s_1^k)\right] \leq \sum_{k,h}\left(4\beta_h^k + \xi_h^k + \zeta_h^k\right)\\ \leq& \sum_{h,s\in\mathcal{S}_h,a\in\mathcal{A}_h,b\in\mathcal{B}_h} c\cdot\sum_{t=1}^{N_h^K(s,a,b)}\sqrt{\frac{H^2S\iota}{t}} + \mathcal{O}(\sqrt{HT\iota})\\ =& \sum_{h,s\in\mathcal{S}_h,a\in\mathcal{A}_h,b\in\mathcal{B}_h}\mathcal{O}\left(\sqrt{H^2S\iota\cdot N_h^K(s,a,b)}\right) + \mathcal{O}(\sqrt{HT\iota})\\ \leq& \sum_{h\in[H]}\mathcal{O}\left(\sqrt{H^2S^2A_hB_hK\iota}\right) \leq \mathcal{O}\left(\sqrt{H^4S^2\left[\max_h A_hB_h\right]K\iota}\right) = \mathcal{O}\left(\sqrt{H^3S^2\left[\max_h A_hB_h\right]T\iota}\right).\end{aligned}$$

This finishes the proof. □

A.2. Proof of Corollary 3

The proof is based on a standard online-to-batch conversion (e.g. (Section 3.1, Jin et al., 2018).) Let $(\widehat{\mu}^k, \widehat{\nu}^k)$ denote the policies deployed by the VI-ULCB algorithm in episode k. We sample $\widehat{\mu}, \widehat{\nu}$ uniformly as

$$\widehat{\mu} \sim \mathrm{Unif}\{\mu^1,\ldots,\mu^K\} \quad \text{and} \quad \widehat{\nu} \sim \mathrm{Unif}\{\nu^1,\ldots,\nu^K\}.$$

Taking expectation with respect to this sampling gives

$$\begin{aligned}&\mathbb{E}_{\widehat{\mu},\widehat{\nu}}\left[V^{\dagger,\widehat{\nu}}(s_1) - V^{\widehat{\mu},\dagger}(s_1)\right] = \frac{1}{K}\sum_{k=1}^K\left[V^{\dagger,\nu^k}(s_1) - V^{\mu^k,\dagger}(s_1)\right]\\ =& \frac{1}{K}\mathrm{Regret}(K) \leq \tilde{\mathcal{O}}\left(\frac{\sqrt{H^3S^2ABT}}{K}\right) \leq \tilde{\mathcal{O}}\left(\sqrt{\frac{H^4S^2AB}{K}}\right),\end{aligned}$$

where we have applied Theorem 2 to bound the regret with high probability. Choosing $K \geq \tilde{\mathcal{O}}(H^4S^2AB/\epsilon^2)$, the right hand side is upper bounded by ϵ, which finishes the proof.

B. Proofs for Section 4

In this section, we prove Theorem 5 and Corollary 6 based on the following lemma about subroutine REWARD_FREE_EXPLORATION. We will defer the proof of this Lemma to Appendix D.

Lemma 3. *Under the preconditions of Theorem 5, with probability at least $1-p$, for any policy μ, ν, we have:*

$$|\hat{V}_1^{\mu,\nu}(s_1) - V_1^{\mu,\nu}(s_1)| \leq \epsilon/2 \tag{3}$$

where $\hat{V}, V$ are the value functions of $\mathrm{MG}(\hat{\mathbb{P}}, \hat{r})$ and $\mathrm{MG}(\mathbb{P}, r)$.

B.1. Proof of Theorem 5

Since both $\inf$ and $\sup$ are contractive maps, by Lemma 3, we have:

$$|\inf_{\nu} V_1^{\hat{\mu},\nu}(s_1) - \inf_{\nu} \hat{V}_1^{\hat{\mu},\nu}(s_1)| \le \epsilon/2$$
$$|\sup_{\mu} V_1^{\mu,\hat{\nu}}(s_1) - \sup_{\mu} \hat{V}_1^{\mu,\hat{\nu}}(s_1)| \le \epsilon/2$$

Since $(\hat{\mu}, \hat{\nu})$ are the Nash Equilibria for $\mathrm{MG}(\hat{\mathbb{P}}, \hat{r})$, we have $\inf_\nu \hat{V}_1^{\hat{\mu},\nu}(s_1) = \sup_\mu \hat{V}_1^{\mu,\hat{\nu}}(s_1)$. This gives:

$$\begin{aligned}\sup_{\mu} V_1^{\mu,\hat{\nu}}(s_1) - \inf_{\nu} V_1^{\hat{\mu},\nu}(s_1) \le& |\sup_{\mu} V_1^{\mu,\hat{\nu}}(s_1) - \sup_{\mu} \hat{V}_1^{\mu,\hat{\nu}}(s_1)| + |\sup_{\mu} \hat{V}_1^{\mu,\hat{\nu}}(s_1) - \inf_{\nu} \hat{V}_1^{\hat{\mu},\nu}(s_1)| \\ &+ |\inf_{\nu} \hat{V}_1^{\hat{\mu},\nu}(s_1) - \inf_{\nu} V_1^{\hat{\mu},\nu}(s_1)| \le \epsilon.\end{aligned}$$

which finishes the proof.

B.2. Proof of Corollary 6

Recall that Theorem 5 requires $T_0 = c(H^5S^2AB\iota/\epsilon^2 + H^7S^4AB\iota^3/\epsilon)$ episodes to obtain an ϵ-optimal policies in the sense:

$$\sup_{\mu} V_1^{\mu,\hat{\nu}}(s_1) - \inf_{\nu} V_1^{\hat{\mu},\nu}(s_1) \le \epsilon.$$

Therefore, if the agent plays the Markov game for T episodes, it can use first T_0 episodes to explore to find ϵ-optimal policies $(\hat{\mu}, \hat{\nu})$, and use the remaining $T - T_0$ episodes to exploit (always play $(\hat{\mu}, \hat{\nu})$). Then, the total regret will be upper bounded by:

$$\mathrm{Regret}(K) \le T_0 \times 1 + (T - T_0) \times \epsilon$$

Finally, choose

$$\epsilon = \max\left\{\left(\frac{H^5S^2AB\iota}{T}\right)^{\frac{1}{3}}, \left(\frac{H^7S^4AB\iota^3}{T}\right)^{\frac{1}{2}}\right\}$$

we finishes the proof.

C. Proofs for Section 5

C.1. Proof of Theorem 9

The theorem is almost an immediate consequence of the general result on mirror descent (Rakhlin & Sridharan, 2013). However, for completeness, we provide a self-contained proof here. The main ingredient in our proof is to show that a "natural" loss estimator satisfies desirable properties—such as unbiasedness and bounded variance—for the standard analysis of mirror descent type algorithms to go through.

Special case of $S = 1$ We first deal with the case of $S = 1$. As the game only has one step ($H = 1$), it reduces to a zero-sum matrix game with a noisy bandit feedback, i.e. there is an unknown payoff matrix $\mathbf{R} \in [0,1]^{A\times B}$, the algorithm plays policies $(\mu_k, \nu_k) \in \Delta_A \times \Delta_B$, observes feedback $r(a^k, b^k) = \mathbf{R}_{a^k,b^k}$ where $(a^k, b^k) \sim \mu_k \times \nu_k$, and the weak regret has form

$$\mathrm{WeakRegret}(T) = \max_{\mu} \sum_{k=1}^{T} \mu^\top \mathbf{R}\nu_k - \min_{\nu} \sum_{k=1}^{K} \mu_k^\top \mathbf{R}\nu.$$

Note that this regret can be decomposed as

$$\mathrm{WeakRegret}(T) = \underbrace{\max_{\mu} \sum_{k=1}^{T} \mu^\top \mathbf{R}\nu_k - \sum_{k=1}^{T} \mu_k^\top \mathbf{R}\nu_k}_{\mathrm{I}} + \underbrace{\sum_{k=1}^{T} \mu_k^\top \mathbf{R}\nu_k - \min_{\nu} \sum_{k=1}^{T} \mu_k^\top \mathbf{R}\nu}_{\mathrm{II}}.$$

We now describe the mirror descent algorithm for the max-player and show that it achieves bound $\mathrm{I} \le \tilde{\mathcal{O}}(\sqrt{AT})$ regardless of the strategy of the min-player. A similar argument on the min-player will yield a regret bound $\mathrm{II} \le \tilde{\mathcal{O}}(\sqrt{BT})$ on the second part of the above regret and thus show $\mathrm{WeakRegret}(T) \le \tilde{\mathcal{O}}(\sqrt{(A+B)T})$.

For all $k \in [T]$, define the loss vector $\ell_k \in \mathbb{R}^A$ for the max-player as

$$\ell_k(a) := e_a^\top \mathbf{R} \nu_k, \quad \text{for all } a \in \mathcal{A}.$$

With this definition the regret I can be written as

$$\mathrm{I} = \max_a \sum_{k=1}^T \ell_k(a) - \sum_{k=1}^T \mu_k(a) \ell_k(a).$$

Now, define the loss estimate $\widetilde{\ell}_k(a)$ as

$$\widetilde{\ell}_k(a) := 1 - \frac{\mathbf{1}\{a^k = a\}}{\mu_k(a)} \big[1 - r(a, b^k)\big].$$

We now show that this loss estimate satisfies the following properties:

(1) Computable: the reward $r(a, b^k)$ is seen when $a = a^k$, and the loss estimate is equal to 1 for all $a \neq a^k$.

(2) Bounded: we have $\widetilde{\ell}_k(a) \le 1$ almost surely for all k and a.

(3) Unbiased estimate of $\ell_k(a)$. For any fixed state $a \in \mathcal{A}$, we have

$$\begin{aligned} &\mathbb{E}\Big[\widetilde{\ell}_k(a) | \mathcal{F}_{k-1}\Big] = 1 - \mu_k(a) \cdot \frac{1}{\mu_k(a)} \mathbb{E}_{b^k \sim \nu_k}\big[1 - r(a, b^k)\big] \\ &= 1 - \big(1 - \mathbb{E}_{b^k \sim \nu_k}[r(a, b^k)]\big) = E_{b^k \sim \nu_k}[r(a, b^k)] = e_a^\top \mathbf{R} \nu_k = \ell_k(a). \end{aligned}$$

(4) Bounded variance: one can check that

$$\begin{aligned} &\mathbb{E}\left[\sum_{a \in \mathcal{A}} \mu_k(a) \widetilde{\ell}_k(a)^2 \big| \mathcal{F}_{k-1}\right] \\ &= \mathbb{E}_{b^k \sim \nu_k}\left[\sum_{a \in \mathcal{A}} \mu_k(a)\big(1 - 2\big(1 - r(a, b^k)\big)\big) + \sum_{a \in \mathcal{A}} (1 - r(a, b^k))^2\right]. \end{aligned}$$

Letting $y_a := 1 - r(a, b^k)$, we have $y_a \in [0, 1]$ almost surely (though it is random), and thus

$$\sum_a \mu_k(a)(1 - 2y_a) + \sum_a y_a^2 \le 1 - 2 \min_a y_a + \sum_a y_a^2 = \sum_{a \neq a_*} y_a^2 + (y_{a^\star} - 1)^2 \le A,$$

where $a^\star = \arg\min_{a \in \mathcal{A}} y_a$.

Therefore, adapting the proof of standard regret-based bounds for the mirror descent (EXP3) algorithm (e.g. (Lattimore & Szepesvári, 2018, Theorem 11.1)), using the loss estimate $\widetilde{\ell}_k(a)$ and taking the step-size to be $\eta_+ \equiv \sqrt{\log A / AT}$, we have the regret bound

$$\mathrm{WeakRegret}_+ \le C \cdot \sqrt{AT \log A},$$

where $C > 0$ is an absolute constant. This shows the desired bound $\tilde{\mathcal{O}}(\sqrt{AT})$ for term I in the regret, and a similar bound $\tilde{\mathcal{O}}(\sqrt{BT})$ holds for term II by using the same algorithm on the min-player.

Algorithm 1 Mirror descent for one-step turn-based game

input Learning rate schedule $(\eta_{+,k}(s), \eta_{-,k}(s))$.
Initialize: Set (μ, ν) to be uniform: $\mu(a|s_1) = \frac{1}{A}$ for all (s_1, a) and $\nu(b|s_2) = \frac{1}{B}$ for all (s_2, b).
for episode $k = 1, \ldots, K$ **do**
Receive s_1.
Play action $a \sim \mu(\cdot|s_1)$. Observe reward $r_1(s_1, a)$ and next state s_2.
Play action $b \sim \nu(\cdot|s_2)$. Observe reward $r_2(s_2, b)$.
Compute $\left\{\widetilde{Q}_1^k(s_1^k, a)\right\}_{a\in\mathcal{A}}$ according to (4) and update

$$\mu^{k+1}(a|s_1^k) \propto \mu^k(a|s_1^k) \cdot \exp(\eta_{+,k}(s_1^k)\widetilde{Q}_1^k(s_1^k, a)).$$

Compute $\left\{\widetilde{Q}_2^k(s_2^k, b)\right\}_{b\in\mathcal{B}}$ according to (5) and update

$$\nu^{k+1}(b|s_2^k) \propto \nu^k(b|s_2^k) \cdot \exp(-\eta_{-,k}(s_2^k)\widetilde{Q}_2^k(s_2^k, b)).$$

end for

Case of $S > 1$ The case of $S > 1$ can be viewed as S independent zero-sum matrix games. We can let both players play the each matrix game independently using an adaptive step-size sequence (such as the EXP3++ algorithm of Seldin & Slivkins (2014)) so that on the game with initial state $s \in \mathcal{S}$ they achieve regret bound

$$\tilde{\mathcal{O}}(\sqrt{(A+B)T_s}),$$

where T_s denotes the number of games that has context s. Summing the above over $s \in \mathcal{S}$ gives the regret bound

$$\mathrm{WeakRegret}(T) \le \sum_s \tilde{\mathcal{O}}(\sqrt{(A+B)T_s}) \le \tilde{\mathcal{O}}(\sqrt{S(A+B)T}),$$

as $\sum_s T_s = T$ and thus $\sum_s \sqrt{T_s} \le \sqrt{ST}$ by Cauchy-Schwarz.

□

C.2. Proof of Theorem 10

We first describe our algorithm for one-step turn-based games ($H = 2$.) Note that this is not equivalent to a zero-sum matrix game, as there is an unknown transition dynamics involved.

As both the max and min player only have one turn to play: $\mu = \{\mu_1\}$ and $\nu = \{\nu_2\}$, in this section we will abuse notation slightly and use (μ, ν) to denote (μ_1, ν_2). We will also use $(\mathcal{A}, \mathcal{B})$ to denote $(\mathcal{A}_1, \mathcal{B}_2)$ for similar reasons.

We now present our mirror descent based algorithm for one-step turn-based games. Define the loss estimates

$$\widetilde{Q}_1^k(s_1^k, a) := 2 - \frac{\mathbf{1}\{a^k = a\}}{\mu^k(a|s_1^k)} \cdot \left[2 - (r(s_1^k, a) + r(s_2^k, b^k))\right] \quad \text{for all } a \in \mathcal{A}, \tag{4}$$

$$\widetilde{Q}_2^k(s_2^k, b) := 1 - \frac{\mathbf{1}\{b^k = b\}}{\nu^k(b|s_2^k)} \cdot \left[1 - r(s_2^k, b)\right] \quad \text{for all } b \in \mathcal{B}. \tag{5}$$

The full algorithm is described in Algorithm 1.

We are now in position to prove the theorem.

Proof of Theorem 10 We begin by decomposing the weak regret into two parts:

$$\begin{aligned}\mathrm{WeakRegret}(T) &= \max_{\mu}\sum_{k=1}^{K} V_1^{\mu,\nu^k}(s_1^k) - \min_{\nu}\sum_{k=1}^{K} V_1^{\mu^k,\nu}(s_1^k)\\ &= \underbrace{\max_{\mu}\sum_{k=1}^{K} V_1^{\mu,\nu^k}(s_1^k) - \sum_{k=1}^{K} V_1^{\mu^k,\nu^k}(s_1^k)}_{\mathrm{WeakRegret}_+} + \underbrace{\sum_{k=1}^{K} V_1^{\mu^k,\nu^k}(s_1^k) - \min_{\nu}\sum_{k=1}^{K} V_1^{\mu^k,\nu}(s_1^k)}_{\mathrm{WeakRegret}_-}.\end{aligned}$$

In the following, we show that both $\mathrm{WeakRegret}_+ \le \mathcal{O}(\sqrt{SAT\iota})$ and $\mathrm{WeakRegret}_- \le \mathcal{O}(\sqrt{SBT\iota})$, which when combined gives the desired result.

Bounding $\mathrm{WeakRegret}_+$ We first consider the case that the initial state is fixed, i.e. $s_1^k \equiv s_1$ for some fixed $s_1 \in \mathcal{S}_1$ and all k. In this case, we have for any μ that

$$V_1^{\mu,\nu^k}(s_1) = \sum_{a\in\mathcal{A}} \mu(a|s_1) Q_1^{\mu,\nu^k}(s_1,a) = \left\langle Q_1^{\mu,\nu^k}(s,\cdot), \mu(\cdot|s_1)\right\rangle_a = \left\langle Q_1^{\cdot,\nu^k}(s_1,\cdot), \mu(\cdot|s_1)\right\rangle_a.$$

Above, the last equality follows by the fact the max player will not play again after the initial action in one-step games, i.e. $Q_1^{\mu,\nu}(s,a)$ does not depend on μ. Applying the above expression, $\mathrm{WeakRegret}_+$ can be rewritten as

$$\mathrm{WeakRegret}_+ = \max_{\mu}\sum_{k=1}^{K}\left\langle Q_1^{\cdot,\nu^k}(s_1,\cdot), \mu(\cdot|s_1)\right\rangle_a - \sum_{k=1}^{K}\left\langle Q_1^{\cdot,\nu^k}(s_1,\cdot), \mu^k(\cdot|s_1)\right\rangle_a,$$

Therefore, bounding $\mathrm{WeakRegret}_+$ reduces to solving an online linear optimization problem over $\Delta_{\mathcal{A}}$ with bandit feedback, where at each step we play μ^k and then suffer a linear loss with loss vector $\left\{Q_1^{\cdot,\nu^k}(s_1,\cdot)\right\}_{a\in\mathcal{A}}$.

Now, recall that our loss estimate in (4), adapted to the setting that $s_1^k \equiv s_1$ can be written as:

$$\widetilde{Q}_1^k(s_1,a) = 2 - \frac{\mathbf{1}\left\{a^k = a\right\}}{\mu^k(a|s_1)}\cdot\left[2 - (r(s_1,a) + r(s_2^k, b^k))\right].$$

We now show that this loss estimate satisfies the following properties:

(1) Computable: the reward $r(s_1,a)$ is seen when $a = a^k$, and the loss estimate is equal to 2 for all other $a \neq a^k$.

(2) Bounded: we have $\widetilde{Q}_1^k(s_1,a) \le 2$ for all k and a.

(3) Unbiased estimate of $Q_1^{\cdot,\nu^k}(s_1,\cdot)$. For any fixed state a, when $a^k = a$ happens, s_2^k is drawn from the MDP transition $\mathbb{P}_1(\cdot|s_1,a)$. Therefore, letting $\mathcal{F}_{k-1}$ be the σ-algebra that encodes all the information observed at the end of episode $k-1$, we have that

$$\widetilde{Q}_1^k(s_1,a)|\mathcal{F}_{k-1} \overset{d}{=} 2 - \frac{\mathbf{1}\left\{a^k = a\right\}}{\mu^k(a|s_1)}\cdot\left[2 - r(s_1,a) - r(s_2^{(a)}, b^{(a)})\right],$$

where $\overset{d}{=}$ denotes equal in distribution, $s_2^{(a)} \sim \mathbb{P}_1(\cdot|s_1,a)$ is an "imaginary" state had we played action a at step 1, and $b^{(a)} \sim \nu^k(\cdot|s_2^{(a)})$. Therefore we have

$$\begin{aligned}&\mathbb{E}\left[\widetilde{Q}_1^k(s_1,a)\middle|\mathcal{F}_{k-1}\right]\\ &= \mathbb{E}_{a\sim\mu^k(\cdot|s_1)}\left[2 - \frac{\mathbf{1}\left\{a = a\right\}}{\mu^k(a|s_1)}\mathbb{E}_{s_2^{(a)},b^{(a)}}\left[2 - r(s_1,a) - r(s_2^{(a)},b^{(a)})\right]\right]\\ &= \mathbb{E}_{s_2^{(a)},b^{(a)}}\left[2 - \frac{\mu^k(a|s_1)}{\mu^k(a|s_1)}\cdot\left[2 - (r(s_1,a) + r(s_2^{(a)},b^{(a)}))\right]\right]\\ &= \mathbb{E}_{s_2^{(a)},b^{(a)}}[r(s_1,a) + r(s_2^{(a)},b^{(a)})] = Q_1^{\cdot,\nu^k}(s_1,a).\end{aligned}$$

(4) Bounded variance: one can check that

$$\begin{aligned}
&\mathbb{E}\left[\sum_{a\in\mathcal{A}}\mu^k(a|s_1)\widetilde{Q}_1^k(s_1,a)^2\big|\mathcal{F}_{k-1}\right]\\
&=4\sum_{a\in\mathcal{A}}\mu^k(a|s_1)\Big(1-\mathbb{E}_{s_2^{(a)},b^{(a)}}[2-r(s_1,a)-r(s_2^{(a)},b^{(a)})]\Big)\\
&\quad+\sum_{a\in\mathcal{A}}\mathbb{E}_{s_2^{(a)},b^{(a)}}[(2-r(s_1,a)-r(s_2^{(a)},b^{(a)}))^2]
\end{aligned}$$

Letting $p_a:=\mu^K(a|s_1)$ and $y_a:=2-r(s_1,a)-r(s_2^{(a)},b_2^{(a)})$, we have $y_a\in[0,2]$ almost surely (though it is random), and thus

$$4\sum_a p_a(1-y_a)+\sum_a y_a^2\le 4(1-\min_a y_a)+\sum_a y_a^2=\sum_{a\neq a_*}y_a^2+(y_{a*}-2)^2\le 4A,$$

where $a*=\arg\min_{a\in\mathcal{A}}y_a$.

Therefore, adapting the proof of standard regret-based bounds for the mirror descent (EXP3) algorithm (e.g. (Lattimore & Szepesvári, 2018, Theorem 11.1)), taking $\eta_+\equiv\sqrt{\log A/AT}$, we have the regret bound

$$\mathrm{WeakRegret}_+\le C\cdot\sqrt{AT\log A},$$

where $C>0$ is an absolute constant.

In the general case where s_1^k are not fixed and can be (in the worst case) adversarial, the design of Algorithm 1 guarantees that for any $s\in\mathcal{S}$, $\mu(\cdot|s)$ gets updated after the k-th episode only if $s_1^k=s$; otherwise the $\mu(\cdot|s)$ is left unchanged. Therefore, the algorithm behaves like solving S bandit problems independently, so we can sum up all the one-state regret bounds of the above form and obtain that

$$\mathrm{WeakRegret}_+\le\sum_{s\in\mathcal{S}}C\sqrt{AT_s\log A}\overset{(i)}{\le}C\sqrt{SAT\log A}=\mathcal{O}(\sqrt{SAT\iota}).$$

where $T_s:=\#\{k:s_1^k=s\}$ denotes the number of occurrences of s among all the initial states, and (i) uses that $\sum_s T_s=T$ and the Cauchy-Schwarz inequality (or pigeonhole principle). Note that we does not know $\{T_s\}_{s\in\mathcal{S}}$ before the algorithm starts to play and thus cannot use $\eta_+(s)=\sqrt{\log A/AT_s}$. We instead use the EXP3++ algorithm (Seldin & Slivkins, 2014) whose step-size $\eta_{+,k}(s)=\sqrt{\log A/AN_k(s)}$ is computable at each episode k.

Bounding $\mathrm{WeakRegret}_-$ For any ν define $r(s_2,\nu(s_2)):=\mathbb{E}_{b\sim\nu(\cdot|s_2)}[r(s_2,b)]$ for convenience. We have

$$\begin{aligned}
&\mathrm{WeakRegret}_-=\sum_{k=1}^K V^{\mu^k,\nu^k}(s_1^k)-\min_\nu\sum_{k=1}^K V^{\mu^k,\nu}(s_1^k)\\
&=\sum_{k=1}^K\mathbb{E}_{a\sim\mu^k(\cdot|s_1)}\big[r(s_1^k,a)+\mathbb{P}_1[r(s_2,\nu^k(s_2))](s_1^k,a)\big]\\
&\quad-\min_\nu\sum_{k=1}^K\mathbb{E}_{a\sim\mu^k(\cdot|s_1)}\big[r(s_1^k,a)+\mathbb{P}_1[r(s_2,\nu(s_2))](s_1^k,a)\big]\\
&\overset{(i)}{=}\mathbb{E}_{a\sim\mu^k(\cdot|s_1)}\left[\sum_{k=1}^K r(s_1^k,a)+\mathbb{P}_1[r(s_2,\nu^k(s_2))](s_1^k,a)\right]\\
&\quad-\sum_{k=1}^K\mathbb{E}_{a\sim\mu^k(\cdot|s_1)}\big[r(s_1^k,a)+\mathbb{P}_1[r(s_2,\nu^\star(s_2))](s_1^k,a)\big]\\
&=\sum_{k=1}^K\mathbb{E}_{a\sim\mu^k(\cdot|s_1),s_2\sim\mathbb{P}_1(\cdot|s_1^k,a)}[r(s_2,\nu^k(s_2))-r(s_2,\nu^\star(s_2))],
\end{aligned}$$

where (i) follows from the fact that if we define $\nu^\star(s_2) = \arg\min_{b'} r(s_2, b')$, then $\nu^\star$ is optimal at every state s_2 and thus also attains the minimum outside. Defining $f_k(s_2) = r(s_2, \nu^k(s_2)) - r(s_2, \nu^\star(s_2))$, we have that $f_k(s_2) \in [0, 1]$ and is a fixed function of s_2 before playing episode k. Thus, if we define

$$\Delta_k = \mathbb{E}_{a,s_2}[f_k(s_2)] - f_k(s_2^k),$$

then Δ_k is a bounded martingale difference sequence adapted to $\mathcal{F}_{k-1}$, so by the Azuma-Hoeffding inequality we have with probability at least $1 - \delta$ that

$$\left|\sum_{k=1}^{K} \Delta_k\right| \le C\sqrt{K\log(1/\delta)} = C\sqrt{T\log(1/\delta)}.$$

On this event, we have

$$\begin{aligned}\text{WeakRegret}_- &= \sum_{k=1}^{K} f_k(s_2^k) + \sum_{k=1}^{K} \Delta_k \\ &\le \underbrace{\sum_{k=1}^{K} \left[r(s_2^k, \nu^k(s_2)) - r(s_2, \nu^\star(s_2))\right]}_{\text{I}} + C\sqrt{K\log(1/\delta)}.\end{aligned}$$

The first term above is the regret for the contextual bandit problem (with context s_2) that the min player faces. Further, the min player in Algorithm 1 plays the mirror descent (EXP3) algorithm independently for each context s_2. Therefore, by standard regret bounds for mirror descent (e.g. Theorem 11.1, (Lattimore & Szepesvári, 2018)) we have (choosing $\eta_- \equiv \sqrt{\log B/T}$ in the fixed s_2 case, and using the EXP3++ scheduling (Seldin & Slivkins, 2014)) for the contextual case), we have

$$\text{I} \le \sum_{s \in \mathcal{S}_2} C\sqrt{BT_s \log B} \le C\sqrt{SBT\log B},$$

which combined with the above bound gives that with high probability

$$\text{WeakRegret}_- \le \mathcal{O}(\sqrt{SBT\iota}),$$

where $\iota = \log(SABT/\delta)$. □

D. Subroutine REWARD_FREE_EXPLORATION

In this section, we present the REWARD_FREE_EXPLORATION algorithm, as well as the proofs for Lemma 3. The algorithm and results presented in this section is simple adaptation of the algorithm in Jin et al. (2020), which studies reward-free exploration in the single-agent MDP setting.

Since the guarantee of Lemma 3 only involves the evaluation of the value under fixed policies, it does not matter whether players try to maximize the reward or minimize the reward. Therefore, to prove Lemma 3 in this section, with out loss of generality, we will treat this Markov game as a single player MDP, where the agent take control of both players' actions in MG. For simplicity, prove for the case $\mathcal{S}_1 = \mathcal{S}_2 = \cdots = \mathcal{S}_H$, $\mathcal{A}_1 = \mathcal{A}_2 = \cdots = \mathcal{A}_H = \mathcal{A}$. It is straightforward to extend the proofs in this section to the setting where those sets are not equal.

The algorithm is described in Algorithm 2, which consists of three loops. The first loop computes a set of policies Ψ. By uniformly sampling policy within set Ψ, one is guaranteed visit all "significant" states with reasonable probabilities. The second loop simply collecting data from such sampling procedure for N episodes. The third loop computes empirical transition and empirical reward by averaging the observation data collected in the second loop. We note Algorithm 2 use subroutine EULER, which is the algorithm presented in Zanette & Brunskill (2019).

We can prove the following lemma, where Lemma 3 is a direct consequence of Lemma 4.

Algorithm 2 REWARD_FREE_EXPLORATION

Input: iteration number N_0, N.
set policy class $\Psi \leftarrow \emptyset$, and dataset $\mathcal{D} \leftarrow \emptyset$.
for all $(s,h) \in \mathcal{S} \times [H]$ **do**
  $r_{h'}(s',a') \leftarrow \mathbb{1}[s'=s \text{ and } h'=h]$ for all $(s',a',h') \in \mathcal{S} \times \mathcal{A} \times [H]$.
  $\Phi^{(s,h)} \leftarrow \text{EULER}(r, N_0)$.
  $\pi_h(\cdot|s) \leftarrow \text{Uniform}(\mathcal{A})$ for all $\pi \in \Phi^{(s,h)}$.
  $\Psi \leftarrow \Psi \cup \Phi^{(s,h)}$.
end for
for $n = 1 \dots N$ **do**
  sample policy $\pi \sim \text{Uniform}(\Psi)$.
  play $\mathcal{M}$ using policy π, and observe the trajectory $z_n = (s_1, a_1, r_1, \dots, s_H, a_H, r_H, s_{H+1})$.
  $\mathcal{D} \leftarrow \mathcal{D} \cup \{z_n\}$
end for
for all $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$ **do**
  $N_h(s,a) \leftarrow \sum_{(s_h,a_h) \in \mathcal{D}} \mathbb{1}[s_h = s, a_h = a]$.
  $R_h(s,a) \leftarrow \sum_{(s_h,a_h,r_h) \in \mathcal{D}} r_h \mathbb{1}[s_h = s, a_h = a]$.
  $\hat{r}_h(s,a) \leftarrow R_h(s,a)/N_h(s,a)$.
  for all $s' \in \mathcal{S}$ **do**
    $N_h(s,a,s') \leftarrow \sum_{(s_h,a_h,s_{h+1}) \in \mathcal{D}} \mathbb{1}[s_h = s, a_h = a, s_{h+1} = s']$.
    $\hat{\mathbb{P}}_h(s'|s,a) \leftarrow N_h(s,a,s')/N_h(s,a)$.
  end for
end for
Return: empirical transition $\hat{\mathbb{P}}$, empirical reward $\hat{r}$.

Lemma 4. *There exists absolute constant $c > 0$, for any $\epsilon > 0$, $p \in (0,1)$, if we set $N_0 \geq cS^3AH^6\iota^3/\epsilon$, and $N \geq cH^5S^2A\iota/\epsilon^2$ where $\iota := \log(SAH/(p\epsilon))$, then with probability at least $1-p$, for any policy π:*

$$|\hat{V}_1^{\pi}(s_1) - V_1^{\pi}(s_1)| \leq \epsilon/2$$

where $\hat{V}, V$ are the value functions of $\text{MG}(\hat{\mathbb{P}}, \hat{r})$ and $\text{MG}(\mathbb{P}, r)$, and $(\hat{\mathbb{P}}, \hat{r})$ is the output of the algorithm 2.

Proof. The proof is almost the same as the proof of Lemma 3.6 in Jin et al. (2020) except that there is no error in estimating r in Jin et al. (2020). We note the error introduced by the difference of $\hat{r}$ and r is a same or lower order term compared to the error introduced by the difference of $\hat{\mathbb{P}}$ and $\mathbb{P}$. We can bound the former error using the similar treatment as in bounding the latter error. This finishes the proof. □

E. Connection to Algorithms against Adversarial Opponents and R-MAX

Similar to the standard arguments in online learning, we can use any algorithm with low regret against adversarial opponent in Markov games to design a provable self-play algorithm with low regret.

Formally, suppose algorithm $\mathcal{A}$ has the following property. The max-player runs algorithm $\mathcal{A}$ and has following guarantee:

$$\max_{\mu} \sum_{k=1}^{K} V_1^{\mu,\nu^k}(s_1^k) - \sum_{k=1}^{K} V_1^{\mu^k,\nu^k}(s_1^k) \leq f(S,A,B,T) \tag{6}$$

where $\{\mu_k\}_{k=1}^K$ are strategies played by the max-player, $\{\nu_k\}_{k=1}^K$ are the possibly adversarial strategies played by the opponent, and function f is a regret bound depends on S, A, B, T. Then, by symmetry, we can also let min-player runs the same algorithm $\mathcal{A}$ and obtain following guarantee:

$$\sum_{k=1}^{K} V_1^{\mu^k,\nu^k}(s_1^k) - \min_{\nu} \sum_{k=1}^{K} V_1^{\mu^k,\nu}(s_1^k) \leq f(S,B,A,T).$$

This directly gives a self-play algorithm with following regret guarantee

$$\begin{aligned}&\mathrm{WeakRegret}(T) = \max_{\mu} \sum_{k=1}^{K} V_1^{\mu,\nu^k}(s_1^k) - \min_{\nu} \sum_{k=1}^{K} V_1^{\mu^k,\nu}(s_1^k) \\ &= \max_{\mu} \sum_{k=1}^{K} V_1^{\mu,\nu^k}(s_1^k) - \sum_{k=1}^{K} V_1^{\mu^k,\nu^k}(s_1^k) + \sum_{k=1}^{K} V_1^{\mu^k,\nu^k}(s_1^k) - \min_{\nu} \sum_{k=1}^{K} V_1^{\mu^k,\nu}(s_1^k) \leq f(S,A,B,T) + f(S,B,A,T)\end{aligned}$$

However, we note there are two notable cases, despite they are also results with guarantees against adversarial opponent, their regret are not in the form (6), thus can not be used to give self-play algorithm, and obtain regret bound in our setting.

The first case is R-MAX algorithm (Brafman & Tennenholtz, 2002), which studies Markov games, with guarantees in the following form.

$$\sum_{k=1}^{K} V_1^{\mu^\star,\nu^\star}(s_1^k) - \sum_{k=1}^{K} V_1^{\mu^k,\nu^k}(s_1^k) \leq g(S,A,B,T)$$

where $\{\mu_k\}_{k=1}^K$ are strategies played by the max-player, $\{\nu_k\}_{k=1}^K$ are the adversarial strategies played by the opponent, $(\mu^\star, \nu^\star)$ are the Nash equilibrium of the Markov game, g is a bound depends on S, A, B, T. We note this guarantee is weaker than (6), and thus can not be used to obtain regret bound in the setting of this paper.

The second case is algorithms designed for adversarial MDP (see e.g. Zimin & Neu, 2013; Rosenberg & Mansour, 2019; Jin et al., 2019), whose adversarial opponent can pick adversarial reward function. We note in Markov games, the action of the opponent not only affects the reward received but also affects the transition to the next state. Therefore, these results for adversaril MDP with adversarial rewards do not directly apply to the setting of Markov game.

References

Brafman, R. I. and Tennenholtz, M. R-max-a general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research*, 3(Oct):213–231, 2002.

Jin, C., Allen-Zhu, Z., Bubeck, S., and Jordan, M. I. Is Q-learning provably efficient? In *Advances in Neural Information Processing Systems*, pp. 4868–4878, 2018.

Jin, C., Jin, T., Luo, H., Sra, S., and Yu, T. Learning adversarial markov decision processes with bandit feedback and unknown transition. *arXiv preprint arXiv:1912.01192*, 2019.

Jin, C., Krishnamurthy, A., Simchowitz, M., and Yu, T. Reward-free exploration for reinforcement learning. *arXiv preprint arXiv:2002.02794*, 2020.

Lattimore, T. and Szepesvári, C. Bandit algorithms. 2018.

Rakhlin, S. and Sridharan, K. Optimization, learning, and games with predictable sequences. In *Advances in Neural Information Processing Systems*, pp. 3066–3074, 2013.

Rosenberg, A. and Mansour, Y. Online convex optimization in adversarial markov decision processes. *arXiv preprint arXiv:1905.07773*, 2019.

Seldin, Y. and Slivkins, A. One practical algorithm for both stochastic and adversarial bandits. In *ICML*, pp. 1287–1295, 2014.

Zanette, A. and Brunskill, E. Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. *arXiv preprint arXiv:1901.00210*, 2019.

Zimin, A. and Neu, G. Online learning in episodic markovian decision processes by relative entropy policy search. In *Advances in neural information processing systems*, pp. 1583–1591, 2013.