A Proof of Lemma 1

For any $\pi_i' \in \Pi_i$, we have

$$\begin{aligned} V_{i,1}^{(\pi_i', \pi_{-1})}(s_1) - V_{i,1}^{\pi}(s_1) &\leq \max_{\pi_i' \in \Pi_i} V_{i,1}^{(\pi_i', \pi_{-i})}(s_1) - V_{i,1}^{\pi}(s_i) \\ &\leq Expl(\pi) \leq \epsilon. \end{aligned}$$

This bound holds for all players and thus we finish the proof.

B Proof of Lemma 2

This lemma can be proved by induction on h. Below discussion is based on the event that Proposition 1 holds for all $i \in [N]$, $k \in \mathbb{N}$ and $h \in [H]$.

First we consider $h = H$ and any $k \in \mathbb{N}$. For any $s \notin \mathcal{S}_i$, the choice of π_i has no influence on $V_{i,H}^{\pi_i, \pi_{-i}^k}(s)$ and we always have $\max_{\pi_i} V_{i,H}^{(\pi_i, \pi_{-i}^k)}(s) = V_{i,H}^{\pi^k}(s) \leq V_{i,H}^k(s)$. For any $s \in \mathcal{S}_i$, we first denote that $\pi_i' = \arg\max_{\pi_i} V_{i,H}^{(\pi_i, \pi_{-i}^k)}(s)$. Then we have that $V_{i,H}^{(\pi_i', \pi_{-i}^k)}(s) \leq Q_{i,H}^k(s, \pi_i'(s,H)) \leq V_{i,H}^k(s)$.
Then we assume that for $h+1$, $h \in [H-1]$, any $k \in \mathbb{N}$ and $s \in \mathcal{S}$, we assume that

$$\max_{\pi_i \in \Pi_i} V_{i,h+1}^{(\pi_i, \pi_{-i}^k)}(s) \leq V_{i,h+1}^k(s).$$

Now we turn to depth h. If $s \in \mathcal{S}_i$, we denote $\pi_i' = \arg\max_{\pi_i \in \Pi_i} V_{i,h}^{(\pi_i, \pi_{-i}^k)}(s)$ and $a' = \pi_i'(s,h)$ for convenience. Then we have

$$\begin{aligned} \max_{\pi_i \in \Pi_i} V_{i,h}^{(\pi_i, \pi_{-i}^k)}(s) \leq & r_i(s, a', h) + \sum_{s' \in \mathcal{S}} P(s'|s, a', h) \max_{\pi_i \in \Pi_i} V_{i,h+1}^{(\pi_i, \pi_{-i}^k)}(s') \\ \leq & \bar{r}_i^k(s, a', h) + \bar{P}^k(s, a', h)^\top V_{i,h+1}^k + b_h^k(s, a) \\ \leq & Q_{i,h}^k(s, a') \\ \leq & V_{i,h}^k(s). \end{aligned}$$

The second inequality holds by using Property 1 of ϕ and the induction assumption.
If $s \notin \mathcal{S}_i$, it is simple to have

$$\begin{aligned} & \max_{\pi_i \in \Pi_i} V_{i,h}^{(\pi_i, \pi_{-i}^k)}(s) \\ \leq & r_i(s, \pi^k(s,h), h) + \sum_{s' \in \mathcal{S}} P(s'|s, \pi^k(s,h), h) \max_{\pi_i \in \Pi_i} V_{i,h+1}^{(\pi_i, \pi_{-i}^k)}(s') \\ \leq & \bar{r}_i^k(s, \pi^k(s,h), h) + \bar{P}^k(s, \pi^k(s,h), h)^\top V_{i,h+1}^k + b_h^k(s, \pi^k(s,h)) \\ \leq & V_h^k(s). \end{aligned}$$

The second inequality holds using Proposition 1. Therefore, we finish our proof with induction.

C supplementary results for Section 6.1

C.1 Failure events

For episode k, we denote $w_h^k(s)$ to be the probability of reaching state s at depth h following policy π^k. Also, we denote $n_h^k(s,a)$ denote the count of visiting state-action pair (s,a) at depth h.Then we define some notations for the

proof:

$$
\begin{aligned}
&w_{\min} = \frac{\epsilon}{2H^2S} \\
&\mathcal{E}_h^k = \{x \in \mathcal{S} \times \mathcal{A} : w_h^k(x) \geq w_{\min}\} \\
&\text{llnp}(n) = \ln(\ln(n)).
\end{aligned}
$$

Notice that ϵ only appears in our analysis. Thus our method finds approximate NE for arbitrary $\epsilon > 0$, like UBEV does.

Then we define some failure episodes that happen with a small probability. For $\delta' \in (0,1)$, we define

$$
\begin{aligned}
F_1^k =& \left\{ \exists x \in \mathcal{S} \times \mathcal{A},\ h \in [H] : n^k(x,h) < \frac{1}{2} \sum_{k'<k} w_h^{k'}(x) - \ln(\frac{SAH}{\delta'}) \right\} \\
F_2^k =& \left\{ \exists x \in \mathcal{S} \times \mathcal{A}, h \in [H] :\ ||(\bar{P}^k(x,h) - P(x,h))||_1 > \sqrt{\frac{4}{n^k(x,h)} \left(2\text{llnp}(n^k(x,h)) + \ln\left(\frac{3SAH(2^S-2)}{\delta'}\right)\right)} \right\} \\
F_3^k =& \left\{ \exists x \in \mathcal{S} \times \mathcal{A}, h \in [H], i \in [N] :\ |\bar{r}_i(x,h) - r_i(x,h)| > \sqrt{\frac{1}{n^k(x,h)} \left(2\text{llnp}(n^k(x,h)) + \ln\left(\frac{3SAH}{\delta'}\right)\right)} \right\} \\
F_4^k =& \left\{ \exists x, x' \in \mathcal{S} \times \mathcal{A},\ h \in [H], u < h : n^k(x,h) < \frac{1}{2} n^k(x',u) \sum_{i<k} w_{u,h}^k(x|x') - \ln\left(\frac{S^2A^2H^2}{\delta'}\right) \right\} \\
F_5^k =& \left\{ \exists x \in \mathcal{S} \times \mathcal{A},\ h \in [H], i \in [N] : \left| (\bar{P}_k(x,h) - P(x,h))^\top V_{i,h+1}^{*,k} \right| \geq \sqrt{\frac{(H-h)^2}{n^k(x,h)} \left(2\text{llnp}(n^k(x,h)) + \ln\frac{3SAH}{\delta'}\right)} \right\} \\
F_6^k =& \Bigg\{ \exists x \in \mathcal{S} \times \mathcal{A},\ s' \in \mathcal{S},\ h \in [H] : |\bar{P}^k(s'|x,h) - P(s'|x,h)| \geq \sqrt{\frac{2P(s'|x)}{n^k(x,h)} \left(2\text{llnp}(n^k(x,h)) + \ln\frac{3S^2AH}{\delta'}\right)} \\
&\qquad + \frac{1}{n^k(x,h)} \left(2\text{llnp}(n^k(x,h)) + \ln\frac{3S^2AH}{\delta'}\right) \Bigg\},
\end{aligned}
$$

where $V_{i,h+1}^{*,k}(s) = \max_{\pi_i \in \Pi_i} V_{i,h+1}^{(\pi_i, \pi_{-i}^k)}(s)$ and $w_{u,h}^k(x|x')$ is the probability to reach x at depth h conditioning on x' is reached at depth u following π^k.

Lemma 3. *(Probabilities for failure events) For $\delta' \in (0,1)$, the bellow inequalities hold*

$\mathbb{P}(\cup_{k=1}^\infty F_1^k) \leq \delta'$, $\mathbb{P}(\cup_{k=1}^\infty F_2^k) \leq \delta'$, $\mathbb{P}(\cup_{k=1}^\infty F_3^k) \leq 2N\delta'$, $\mathbb{P}(\cup_{k=1}^\infty F_4^k) \leq \delta'$, $\mathbb{P}(\cup_{k=1}^\infty F_5^k) \leq 2N\delta'$, $\mathbb{P}(\cup_{k=1}^\infty F_6^k) \leq 2\delta'$.

Proof. This lemma is highly relative to the appendices E.2 of UBEV. Below corollaries are all presented in [Dann et al., 2017].

Specifically, F_1^k here is exactly the F_k^N in UBEV. Thus we can get $\mathbb{P}(\cup_{k=1}^\infty F_1^k) \leq \delta'$ by directly apply Corollary E.4.

For F_2^k, we can directly refer to Corollary E.3 and get $\mathbb{P}(\cup_{k=1}^\infty F_2^k) \leq \delta'$.

Our definition of F_3^k is almost the same as F_k^R in UBEV except that we need to bound the immediate rewards for all players now. Thus we can prove $\mathbb{P}(\cup_{k=1}^\infty F_3^k) \leq 2N\delta'$ with Corollary E.1 and a union bound over all players.

Then F_4^k corresponds to F_k^{CN} and $\mathbb{P}(\cup_{k=1}^\infty F_4^k) \leq \delta'$ with Corollary E.4.

F_5^k extends F_k^V to FTSGs. Notice that in MDPs, V_{t+1}^* is a fixed vector, while in FTSGs, this does not holds. Therefore we need carefully consider the relationship of $P(x,h)$ and $V_{i,h+1}^{*,k}$. Since we consider the time-dependent dynamics, we always have $P(x,h)$ and $V_{i,h+1}^{*,k}$ independent for all players. Therefore, we can apply Corollary E.1 in [Dann et al., 2017] to get that $\mathbb{P}(\cup_{k=1}^\infty F_5^k) \leq 2N\delta'$.

Finally for F_6^k, it corresponds to F_k^P and has $\mathbb{P}(\cup_{k=1}^\infty F_6^k) \leq 2\delta'$ with Corollary E.2. □

Then we define the union of all these failure events as

$$F = \cup_k [F_1^k \cup F_2^k \cup F_3^k \cup F_4^k \cup F_5^k \cup F_6^k].$$

We can get the conclusion that $\mathbb{P}(F) \leq \delta$ with lemma 3 and letting $\delta' = \delta/(4N+5)$. That is the supplement set of F, denoted as F^c, happens with a probability at least $1-\delta$.

C.2 Property 1 requirement

We choose the bonus function as

$$\phi^{UPAC} = (H+1)\sqrt{\frac{2\ln\ln(\max\{e, n^k(s,a,h)\}) + \ln((24N+30)SA/\delta)}{n^k(s,a,h)}}.$$

On event F^c, we have that

$$\begin{aligned}
&|(r_i^k(s,a,h) - \bar{r}_i^k(s,a,h)) + \left(P^k(s,a,h) - \bar{P}^k(s,a,h)\right)^\top V_{i,h+1}^{*,k}| \\
\leq &|(r_i^k(s,a,h) - \bar{r}_i^k(s,a,h))| + |\left(P^k(s,a,h) - \bar{P}^k(s,a,h)\right)^\top V_{i,h+1}^{*,k}| \\
\leq &b_h^k(s,a),
\end{aligned}$$

where the last inequality holds using the definition of F_3^k and F_5^k.

Therefore, our design of ϕ^{UPAC} satisfies Property 1.

C.3 Nice and Friendly episodes

Then we define the nice episodes:

Definition 4. *Episode k is called **nice episode** if $\forall\, x \in \mathcal{S}\times\mathcal{A}$, at least one of the below two conditions holds:*

$$(1)\ w_h^k(x) < w_{\min},\ \forall\, h \in [H];$$
$$(2)\ \frac{1}{4}\sum_{i<k}\sum_{h'=1}^{H} w_{h'}^i(x) \geq \ln(\frac{SAH}{\delta'}).$$

Next we define friendly episodes:

Definition 5. *Episode k is called **friendly episode** if $\forall\, x, x' \in \mathcal{S}\times\mathcal{A}$, at least one of the below two conditions holds:*

$$(1)\ w_{u,h}^k(x|x') < w_{\min},\ \forall\, h \in [H];$$
$$(2)\ \frac{1}{4}\sum_{i<k}\sum_{h'=1}^{H} w_{u,h}^k(x|x') \geq H\ln(\frac{S^2A^2H^2}{\delta'}).$$

This is exactly the same as the definition in UBEV.

With Lemma E.2 in [Dann et al., 2017] we can bound the number of episodes that are not nice or friendly on F^c.

Lemma 4. *(Sample complexity for non-nice episodes) On F^c, the number of episodes which are not nice is no more than*

$$\frac{6H^3S^2A}{\epsilon}\ln\left(\frac{HSA}{\delta'}\right).$$

Lemma 5. *(Sample complexity for non-friendly episodes) On F^c, the number of episodes which are not friendly is no more than*

$$\frac{48H^4S^3A^2}{\epsilon}\ln\left(\frac{H^2S^2A^2}{\delta'}\right).$$

Then we can concentrate on nice and friendly episodes.

Lemma 6. *(Property of nice episodes) Let $r \geq 1$. Let $D \geq 1$ be a poly-logarithmic function of relevant parameters. For $\epsilon' > 0$ there are at most*

$$\frac{8H^r SA}{\epsilon'^r}\text{polylog}(S, A, H, \delta'^{-1}, \epsilon'^{-1})$$

nice episodes such that

$$\sum_{h=1}^{H}\sum_{s\in\mathcal{E}_h^k} w_h^k(x)\left(\frac{\text{llnp}(n^k(x,h))+D}{n^k(x,h)}\right)^{1/r} > \epsilon'.$$

This lemma can be directly proved with Lemma E.3 of [Dann et al., 2017].

Lemma 7. *(Property of friendly episodes) On good event F^c, there are at most*

$$(\frac{9216}{\epsilon}+417S)\text{polylog}(S, A, H, \delta'^{-1}, \epsilon'^{-1})$$

friendly episodes such that

$$V_{i,1}^{*,k} - V_{i,1}^{\pi^k} \geq \epsilon,$$

if $\delta' \leq \frac{3AS^2H}{e^2}$

This lemma can be derived from Lemma E.8 of [Dann et al., 2017].

C.4 The Sample Complexity for One Player

Now we give below lemma:

Lemma 8. *Following UBVP, with a probability $1-\delta$, for any $\epsilon > 0$ and any $i \in [N]$, the number of episodes that $\Delta_i^k > \epsilon$ is at most*

$$O\left(\frac{H^4S^2A}{\epsilon^2}\text{polylog}\left(N, H, S, A, \frac{1}{\delta}, \frac{1}{\epsilon}\right)\right).$$

Proof. For arbitrary player $i \in [N]$, we can decompose $\Delta_i^k = V_{i,1}^k(s_1) - V_{i,1}^{\pi^k}(s_1)$ to get

$$\begin{aligned}
&V_{i,1}^k(s_1) - V_{i,1}^{\pi^k}(s_1)\\
=&\bar{r}_i^k(s_1,\pi(s_1,1),1) - r_i(s_1,\pi(s_1,1),1) + \bar{P}^k(s_1,\pi(s_1,1),1)^\top V_{i,2}^k - P(s_1,\pi(s_1,1),1)^\top V_{i,2}^{\pi^k} + b_1^k(s_1,\pi(s_1,1))\\
=&(\bar{r}_i^k(s_1,\pi(s_1,1),1) - r_i(s_1,\pi(s_1,1)),1) + \left(\bar{P}^k(s_1,\pi(s_1,1),1) - P(s_1,\pi(s_1,1),1)\right)^\top V_{i,2}^k\\
&+ P(s_1,\pi(s_1,1),1)^\top\left(V_{i,2}^k - V_{i,2}^{\pi^k}\right) + b_1^k(s_1,\pi(s_1,1))\\
=&\sum_{h=1}^{H}\sum_{x\in\mathcal{S}\times\mathcal{A}} w_h^k(x)\left((\bar{r}_i^k(x,h) - r_i(x,h)) + \left(\bar{P}^k(x,h) - P(x,h)\right)^\top V_{i,h+1}^k + b_h^k(x)\right)\\
=&\sum_{h=1}^{H}\sum_{x\in\mathcal{E}_h^k} w_h^k(x)\left((\bar{r}_i^k(x,h) - r_i(x,h)) + \left(\bar{P}^k(x,h) - P(x,h)\right)^\top V_{i,h+1}^k + b_h^k(x)\right)\\
&+\sum_{h=1}^{H}\sum_{x\notin\mathcal{E}_h^k} w_h^k(x)\left((\bar{r}_i^k(x,h) - r_i(x,h)) + \left(\bar{P}^k(x,h) - P(x)\right)^\top V_{i,h+1}^k + b_h^k(x,h)\right)\\
\leq&\sum_{h=1}^{H}\sum_{x\in\mathcal{E}_h^k} w_h^k(x)\left((\bar{r}_i^k(x,h) - r_i(x,h)) + \left(\bar{P}^k(x,h) - P(x,h)\right)^\top V_{i,h+1}^k + b_h^k(x)\right) + \sum_{h=1}^{H}\sum_{x\notin\mathcal{E}_h^k} Hw_{\min}.
\end{aligned}$$

Recall that we consider deterministic policies here. Thus the second term can be bounded by $H^2Sw_{\min} = \frac{1}{2}\epsilon$.

Thus we just need to upper bound the first term. Recall that

$$b_h^k(x) = (H+1)\sqrt{\frac{1}{n^k(x,h)}\left(2\text{llnp}(n^k(x,h)) + \ln\left(\frac{6SAH}{\delta'}\right)\right)}.$$

We also have that on good event F^c

$$\begin{aligned}
\left(\bar{P}^k(x,h) - P(x,h)\right)^\top V_{i,h+1}^k \leq & ||\bar{P}^k(x,h) - P(x,h)||_1 ||V_{i,h+1}^k||_\infty \\
\leq & \sqrt{\frac{4H^2}{n^k(x,h)}\left(2\text{llnp}(n^k(x,h)) + \ln\left(\frac{3SAH(2^S-2)}{\delta'}\right)\right)} \\
\leq & \sqrt{\frac{4H^2S}{n^k(x)}\left(2\text{llnp}(n^k(x)) + \ln\left(\frac{6SAH}{\delta'}\right)\right)} \\
\bar{r}_i^k(x,h) - r_i(x,h) \leq & \sqrt{\frac{1}{n^k(x,h)}\left(2\text{llnp}(n^k(x,h)) + \ln\left(\frac{3SAH}{\delta'}\right)\right)}.
\end{aligned}$$

Combine them and we get

$$\begin{aligned}
& \sum_{h=1}^H \sum_{x \in \mathcal{E}_h^k} w_h^k(x) \left((\bar{r}_i^k(x,h) - r_i(x,h)) + \left(\bar{P}^k(x,h) - P(x,h)\right)^\top V_{i,h+1}^k + b_h^k(x) \right) \\
\leq & (2H\sqrt{S} + H + 2) \sum_{h=1}^H \sum_{s \in \mathcal{E}_h^k} w_h^k(x) \sqrt{\frac{1}{n^k(x,h)}\left(2\text{llnp}(n^k(x,h)) + \ln\left(\frac{6SAH}{\delta'}\right)\right)}. \qquad (4)
\end{aligned}$$

Then we let $r = 2$, $D = \ln(6SA/\delta')$ and $\epsilon' = \epsilon/(4H\sqrt{S}+2)$. By applying lemma 6, there are at most

$$\frac{32HSA(2H\sqrt{S}+1)^2}{\epsilon^2}\text{polylog}(S, A, H, \delta'^{-1}, \epsilon^{-1})$$

nice episodes such that

$$\sum_{h=1}^H \sum_{x \in \mathcal{E}_h^k} w_h^k(s) \left((\bar{r}^k(x,h) - r(x,h)) + \left(\bar{P}^k(x,h) - P(x,h)\right)^\top V_{h+1}^{k,U} + b_h^{k,U}(x) \right) > \epsilon.$$

Therefore, by choosing $\delta' = \delta/(4N+5)$, we finish our proof. □

C.5 Proof of theorem 1

Here we give the proof of our main theorem. Our target is give the sample complexity of L^ϵ. Since all players are involved in L^ϵ, we solve this problem by bounding each player separately.

Recall our definition for the the *best response distance* of player i for π as

$$Bsd_i(\pi) := \max_{\pi_i' \in \Pi_i} V_{i,1}^{(\pi_i', \pi_{-i})}(s_1) - V_{i,1}^{(\pi)}(s_1).$$

We rewrite L^ϵ with $Bsd_i(\pi^k)$ where $\pi^k = (\pi_1^k, \pi_2^k, ..., \pi_N^k)$ is the policy tuple used during episode k. We now can bound L^ϵ with

$$L^\epsilon \leq \sum_{k \in \mathbb{N}} \mathbb{I}\left[\exists i \in [N], Bsd_i(\pi^k) > \epsilon\right].$$

It is easy to see that in UBVP, players are symmetric, and thus the result for one can be adapted to the others. Now we consider player i, $i \in [N]$. Now we define $\Delta_i^k := V_{i,1}^k(s_1) - V_{i,1}^{\pi^k}(s_1)$. Now we give the key lemma that connect Δ_i^k and Bsd_i. Our design for UBVP exactly ensure such a good connection. Recall Lemma 2, we have that on good events F^c, for any $i \in [N]$, $k \in \mathbb{N}$, $h \in [H]$ and $s \in \mathcal{S}$,

$$\max_{\pi_i \in \Pi_i} V_{i,h}^{(\pi_1, \pi_{-i}^k)}(s) \leq V_{i,h}^k(s).$$

Lemma 2 can lead to the result that on F^c, for all $i \in [N]$,

$$Bsd_i(\pi^k) = \max_{\pi_i \in \Pi_i} V_{i,1}^{(\pi_i, \pi_{-i}^k)}(s_1) - V_{i,1}^{\pi^k}(s_1) \leq \Delta_i^k.$$

Naturally, on good events F^c,

$$L^\epsilon \leq \sum_{i \in [N]} \sum_{k \in \mathbb{N}} \mathbb{I}\left[\Delta_i^k > \epsilon\right].$$

Further we notice that applying lemma 6 to the right hand side of δ_i^k. This process in fact is independent of the choice of players. That is, this holds for all players at the same time. Hence the sample complexity in lemma 8 is exactly the sample complexity of L^ϵ. It might be strange that the number of player N only appears in the polylog term. A rethinking can remind us that the number of states S in fact include the complexity of player numbers implicitly, because S is the number of states for all players.

Therefore we finish the proof of first part of theorem 1.

For the second part of the theorem, since the analysis is relies on the exact values of player i. We cannot removes the term N. Thus from Lemma 7 and Lemma 5, we can get that on good event F^c:

$$L^\epsilon \leq O((NS + \frac{N + H^4 S^3 A^2}{\epsilon})\text{polylog}(N, S, A, H, \delta^{-1}, \epsilon^{-1})).$$

Therefore we finish the proof.

D supplementary results for Section 6.2

Recall that we design two ϕ functions as:

$$\begin{aligned}
\phi_1^{HPR} =& 8HL\sqrt{1/n^k(s,a,h)}, \\
\phi_2^{HPR} =& \sqrt{\frac{8LVar_{s' \sim \bar{P}(s,a,h)} V_{i,h+1}^k(s')}{n^k(s,a,h)}} + \frac{14HL}{3n^k(s,a,h)} + HL\sqrt{1/n^k(s,a,h)} + \sqrt{\frac{8\sum_{s'} \bar{P}(s,a,s',h) C(s')}{n^k(s,a,h)}},
\end{aligned}$$

where $C(s') = \min\{10^4 H^3 S^2 A L^2 / n^k(s,a,s',h), H^2\}$ and $L = \ln(5HSATN/\delta)$.

We add a term $HL\sqrt{1/n^k(s,a,h)}$ to bonus 1 and 2 in [Azar et al., 2017] to get the two bonus functions. We add this extra term because we assume reward functions are random functions. Thus we need extra term to upper bound the gap caused by r_i.

E Experiment Implementations

Here we give detailed implementations for our empirical results. We give implementations of methods and games.

E.1 Methods

With 5 baselines, we implement 6 methods. We list them below.

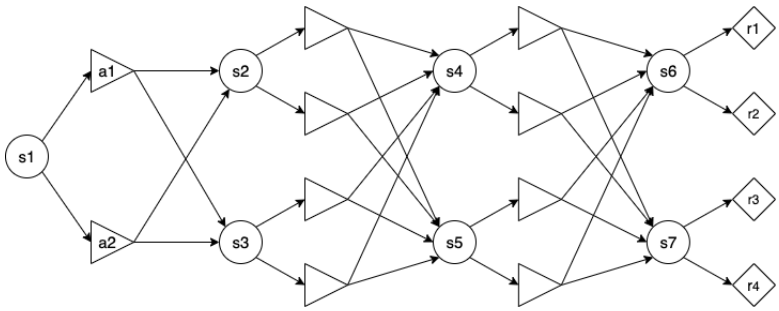

Figure 3: Two-player zero-sum FTSG

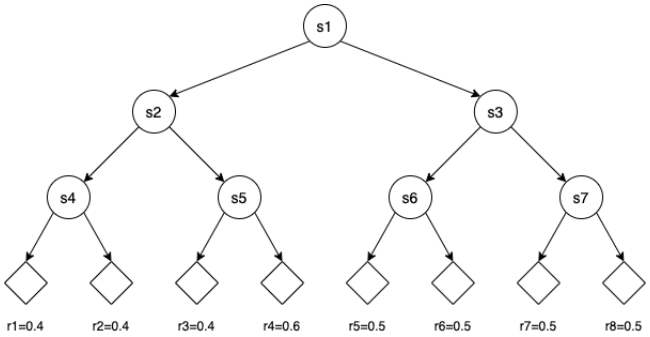

Figure 4: Cooperative FTSG

- UBVP: we implement UBVP with bonus function $\phi_1^H PR$. Considering that our game only gain rewards at terminal nodes, we can refine this bonus function to be $\sqrt{\ln(SAt/\delta)/n^k(s,a,h)}$. We set $\delta = 0.1$.
- MCTS [Coulom, 2006]: we implement MCTS in a way similar to UCT. For each node, the upper bounds for Q values are constructed by the averaged value and a bonus term. We choose the same bonus function as UBVP.
- NashQ [Hu and Wellman, 2003]: for FTSGs, we implement NashQ by letting each player choose its maximal Q values. We use ϵ-greedy as its exploration strategy and we choose $\epsilon = 0.1$.
- CFR-PSRL [Zhou et al., 2020]: this is also a model-based methods which combines the technique of PSRL. In this method, each player has an exploration strategy to interact with the environment. Each game, there is only one player conducts its exploration strategy. We update all policies when each player has explored for 10 times. Notice that we need to use the average policy, so CFR-PSRL define its policy on histeries, rather than states. However, it can maintain the posterior based on states.
- MCCFROS [Lanctot et al., 2009]: similar to ϵ-greedy, MCCFROS also has a probability p to uniform sample actions. We set this p to be 0.1 for all histories.
- FSPFQI [Heinrich et al., 2015]: for FSPFQI we still define policies on states. We keep a deque of size 100 to be the policy pool. We also use ϵ-greedy as the exploration strategy for FSPFQI. We choose $\epsilon = 0.1$.

E.2 Games

Here we give detailed description of our games.

- Two-player zero-sum FTSG: we choose $H = 4$ for this game, as shown in Fig. 3. At depth 1 and 3, player 0 takes actions and at depth 2 and 4, player 1 takes actions. There is only one state for depth 1 and all other 3 depths has 2 states. Therefore this game has 7 different states in total. Each player have two actions, a_1 and a_2 to choose. For depth $h \in [3]$, each state-action pair (s, a) has a non-zero probability to reach either states of depth $h+1$. For

depth 4, after player 1 chooses one action, a reward $R \in \{0, 1\}$ is sampled from a fixed Bernoulli distribution. Player 0 gets R and player 1 gets $-R$.

- Cooperative FTSG: we choose horizon $H = 8$ for this game. There are two players in total and they take actions alternatively. Initially player 0 at state s_0 takes actions. At each state, current player has two actions a_1 and a_2 to choose. The transition of this game is deterministic and the game is expanded as a binary tree. Thus there are $2^8 - 1 = 255$ states in total. The rewards are returned at terminal states. In this game, the two players has the same rewards, so they need to cooperate to get the highest reward. That is, they need to find the SPE solution. The rewards are also sampled from Bernoulli distributions. If at s_0, player 0 chooses a_2, the expected reward for each player will always be 0.5. If player 0 chooses a_1, they are possible to reach 64 terminal nodes and get 128 trajectories. Among the 128 trajectories, there is only one has an expected reward of 0.6, while others have expected reward 0.4. We give an example of $H = 3$ game in Fig. 4. Notice that this game can be solved as an MDP. We use it as an extreme example to show whether algorithms can reach SPEs. Specifically, we considers exploitability of SPEs and thus we change $Expl(T)$ to be $0.6T - \sum_{t=1}^{T} V_{0,1}^t$.
- Three-player FTSG: we choose $H = 6$ for this game. At depth 1 and 4, player 0 takes actions; at depth 2 and 5, player 1 takes actions; at depth 3 and 6, player 2 takes actions. There is only one state for depth 1 and all other 3 depth has 2 states. Therefore this game has 11 different states in total. Each player have two actions, a_1 and a_2 to choose. For depth $h \in [5]$, each state-action pair (s, a) has a non-zero probability to reach either states of depth $h + 1$. For depth h, after player 2 chooses one action, a reward vector $R \in \{0, 1\}^3$ is sampled from three fixed Bernoulli distributions. Each player gets one reward. This game has similar structures as Fig. 3 except that there are 6 depths and terminal state-action pairs return a vector of rewards.