Appendix A Proof of Theorem 4.1

Our proof of regret bounds for Strategic ULCB is very similar to the proof for the regret bounds of Optimistic ULCB [Bai and Jin, 2020]. The key difference is that we need to show that the tighter confidence bounds maintained by Strategic ULCB still constrain the exploitability of the evaluation policies $\tilde{\mu}^k$ and $\tilde{\nu}^k$ (Lemma A.2), and that these confidence bounds still converge under our exploration policies. We also directly bound the L_1 error of the transition model, which leads to a somewhat simpler proof, and helps us better understand the nature of the confidence sets that Strategic ULCB implicitly maintains.

Lemma A.1. *For a given $K \geq 3$ and $\delta > 0$, define β_t as*

$$\beta_t = H\sqrt{\frac{2\left[|S|\ln(KH|S||A||B|/\delta)\right]}{t}} \tag{1}$$

then with probability at least $1-\delta$, for all $k \in [K]$, $h \in H$, $s \in S_h$, $a \in A_{h,s}$ and $b \in B_{h,s}$, and for all $V \in [0, H]^{|S|}$, we have

$$\left|\hat{P}_h^k(s,a,b)^\top V - P(s,a,b)^\top V\right| \leq \beta_t \tag{2}$$

for $t = N_h^k(s,a,b)$.

Proof. When $N_h^k(s,a,b) = 0$, Equation 2 holds trivially as $\beta_t = \infty$. Otherwise, we can apply the well known bound on the L_1 error of an empirical distribution due to Weissman et al. [2003] to show that

$$\Pr\left\{\|\hat{P}_h^k(s,a,b) - P(s,a,b)\|_1 \geq \epsilon\right\} \quad \leq \quad (2^{|S|} - 2)\exp\{-N_h^k(s,a,b)\frac{\epsilon}{2}\} \tag{3}$$

Note that, for all $V \in [0, H]^{|S|}$

$$|\hat{P}_h^k(s,a,b)^\top V - P(s,a,b)^\top V| \quad \leq \quad H\|\hat{P}_h^k(s,a,b) - P(s,a,b)\|_1 \tag{4}$$

and so for $t = N_h^k(s,a,b)$ and β_t defined according to Equation 1 we have

$$\Pr\left\{\exists V, |\hat{P}_h^k(s,a,b)^\top V - P(s,a,b)^\top V| \geq \beta_t\right\} \quad \leq \quad \frac{\delta}{KH|S||A||B|} \tag{5}$$

Taking the union bound over k, h, s, a and b yields the desired result. □

For games with deterministic transitions, $\hat{P}_h^k(s,a,b) = P(s,a,b)$ whenever $N_h^k(s,a,b) > 0$, and so Equation 2 will hold even for $\beta_t = 0$, which is the value we use for our experiments in deterministic games. We can now show that our confidence bounds $\bar{V}_h^k$ and $\underline{V}_h^k$ not only constrain the value of the game at each state, but also bound the exploitability of our evaluation policies $\tilde{\mu}^k$ and $\tilde{\nu}^k$.

Lemma A.2. *When Strategic-ULCB is run with β_t as defined in Equation 1, the for all $k \in [K]$, $h \in H$ and $s \in S_h$, we have*

$$\bar{V}_h^k(s) \geq \sup_{\mu} V_h^{\mu,\nu^k}(s) \tag{6}$$

$$\underline{V}_h^k(s) \leq \inf_{\nu} V_h^{\mu^k,\nu}(s) \tag{7}$$

with probability at least $1-\delta$.

Proof. For each $k \in [K]$ we prove this by induction on h. We will only show the proof for the upper bound, as the proof for the lower bound is symmetric. Assume that for some $h \in [H]$ we have, for all $s \in S_h$

$$\bar{V}_{h+1}^k(s) \geq \sup_{\mu} V_{h+1}^{\mu,\nu^k}(s) \tag{8}$$

By Lemma A.1, Equation 2 will hold simultaneously for all k, h, s, a and b with probability at least $1-\delta$, and so when $N_h^k(s,a,b) > 0$, we have

$$\begin{aligned}
\bar{Q}_h^k(s,a,b) &= \hat{R}_h^k(s,a,b) + \hat{P}_h^k(s,a,b)\bar{V}_{h+1}^k + \beta_t && (9)\\
&\geq R(s,a,b) + \hat{P}_h^k(s,a,b)\bar{V}_{h+1}^k && (10)\\
&\geq R(s,a,b) + P(s,a,b)\sup_{\mu} V_{h+1}^{\mu,\nu^k} && (11)\\
&= \sup_{\mu} Q_h^{\mu,\nu^k}(s,a,b) && (12)
\end{aligned}$$

where the $t = N_h^k(s,a,b)$, and the first inequality also uses the fact that $\hat{R}_h^k(s,a,b) = R(s,a,b)$ when $N_h^k(s,a,b) > 0$. When $N_h^k(s,a,b) = 0$, Equation 9 holds trivially, as $\bar{Q}_h^k(s,a,b) = H$. By the definition of $\bar{V}_{h+1}^k(s)$, we then have

$$\begin{aligned}
\bar{V}_{h+1}^k(s) - \sup_{\mu} V_h^{\mu,\nu^k}(s) &= \mu_h^k(s)^\top \bar{Q}_h^k(s,\cdot,\cdot)\tilde{\nu}_h^k(s) && (13)\\
&\quad - \max_{a \in A_{h,s}} \sup_{\mu} Q_h^{\mu,\nu^k}(s,a,\cdot)\tilde{\nu}_h^k(s) && (14)\\
&\geq \mu_h^k(s)^\top \bar{Q}_h^k(s,\cdot,\cdot)\tilde{\nu}_h^k(s) && (15)\\
&\quad - \max_{a \in A_{h,s}} \sup_{\mu} Q_h^k(s,a,\cdot)\tilde{\nu}_h^k(s) && (16)\\
&= 0 && (17)
\end{aligned}$$

which proves the inductive step. The first inequality follows directly from Equation 9, while the second inequality follows from the fact that $(\mu_h^k(s), \tilde{\nu}_h^k(s))$ for a Nash equilibrium of the matrix game defined by $\bar{Q}_h^k(s,\cdot,\cdot)$, and so $\mu_h^k(s)$ is a best-response to $\tilde{\nu}_h^k(s)$ under $\bar{Q}_h^k(s,\cdot,\cdot)$. Finally, we can see that Equation 8 holds trivially for $h = H+1$, where we implicitly assume that $\bar{V}_h^k(s) = \sup_{\mu} V_h^{\mu,\nu^k}(s) = 0$, which concludes the proof. □

Lemma A.2 will be sufficient to prove Theorem 4.1 and bound the NashConv regret of the evaluation policies $\tilde{\mu}^k$ and $\tilde{\nu}^k$. The remainder of the proof will closely follow the proof for Optimistic ULCB given by Bai and Jin [2020], with slight modifications to account for the presence of separate exploration and evaluation policies.

Proof of Theorem 4.1. We begin with the definition of the NashConv regret

$$\text{Regret}(K) = \sum_{k=1}^{K} \sup_{\mu} V_1^{\mu,\nu^k}(s_1) - \inf_{\nu} V_1^{\mu^k,\nu}(s_1) \tag{18}$$

for any $k \in [K]$ and $h \in [H]$, we have

$$\begin{aligned}
&\sup_{\mu} V_h^{\mu,\nu^k}(s_h^k) - \inf_{\nu} V_h^{\mu^k,\nu}(s_h^k) &(19)\\
&\leq \bar{V}_h^k(s_h^k) - \underline{V}_h^k(s_h^k) &(20)\\
&= \mu_h^k(s_h^k)^\top \bar{Q}_h^k(s_h^k,\cdot,\cdot)\tilde{\nu}_h^k(s_h^k) - \tilde{\mu}_h^k(s_h^k)^\top \underline{Q}_h^k(s_h^k,\cdot,\cdot)\nu_h^k(s_h^k) &(21)\\
&\leq \mu_h^k(s_h^k)^\top \bar{Q}_h^k(s_h^k,\cdot,\cdot)\nu_h^k(s_h^k) - \mu_h^k(s_h^k)^\top \underline{Q}_h^k(s_h^k,\cdot,\cdot)\nu_h^k(s_h^k) &(22)\\
&= \mu_h^k(s_h^k)^\top \left[\bar{Q}_h^k(s_h^k,\cdot,\cdot) - \underline{Q}_h^k(s_h^k,\cdot,\cdot)\right]\nu_h^k(s_h^k) &(23)
\end{aligned}$$

where the first inequality follows from Lemma A.2, while the second follow from the fact that $\tilde{\mu}^k$ and $\tilde{\nu}^k$ are best responses, and so changing to the optimistic strategies μ^k and ν^k can only increase the width of the confidence interval. We can decompose the last term as

$$\begin{aligned}
&\mu_h^k(s_h^k)^\top \left[\bar{Q}_h^k(s_h^k,\cdot,\cdot) - \underline{Q}_h^k(s_h^k,\cdot,\cdot)\right]\nu_h^k(s_h^k) &(24)\\
&= \left[\bar{Q}_h^k - \underline{Q}_h^k\right](s_h^k,a_h^k,b_h^k) + \xi_h^k &(25)\\
&= \hat{P}_h^k(s_h^k,a_h^k,b_h^k)^\top \left[\bar{V}_{h+1}^k - \underline{V}_{h+1}^k\right] + +2\beta_h^k + \xi_h^k &(26)\\
&= P(s_h^k,a_h^k,b_h^k)^\top \left[\bar{V}_h^k - \underline{V}_h^k\right] + 4\beta_h^k + \xi_h^k &(27)\\
&= \left[\bar{V}_{h+1}^k - \underline{V}_{h+1}^k\right](s_{h+1}^k) + \zeta_h^k + 4\beta_h^k + \xi_h^k &(28)
\end{aligned}$$

where $\beta_h^k = \beta_t$ for $t = N_h^k(s,a,b)$. The terms ξ_h^k and ζ_h^k are defined as

$$\begin{aligned}
\xi_h^k &= \mathrm{E}_{a,b\sim\mu_h^k(s_h^k),\nu_h^k(s_h^k)}\left[\bar{Q}_h^k - \underline{Q}_h^k\right](s_h^k,a,b) &(29)\\
&\quad - \left[\bar{Q}_h^k - \underline{Q}_h^k\right](s_h^k,a_h^k,b_h^k) &(30)\\
\zeta_h^k &= \mathrm{E}_{s\sim P(s_h^k,a_h^k,b_h^k)}\left[\bar{V}_{h+1}^k - \underline{V}_{h+1}^k\right](s) &(31)\\
&\quad - \left[\bar{V}_{h+1}^k - \underline{V}_{h+1}^k\right](s_{h+1}^k) &(32)\\
& &(33)
\end{aligned}$$

Here ξ_h^k and ζ_h^k are not i.i.d., but the sequences of their partial sums over k and

h are martingales, and so by the Azuma-Hoeffding inequality

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\xi_h^k \leq \sqrt{2KH^3 \ln\frac{1}{\delta}} \tag{34}$$

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\zeta_h^k \leq \sqrt{2KH^3 \ln\frac{1}{\delta}} \tag{35}$$

we then have

$$\sum_{k=1}^{K} \sup_{\mu} V_1^{\mu,\nu^k}(s_1) - \inf_{\nu} V_1^{\mu^k,\nu}(s_1) \tag{36}$$

$$\leq \sum_{k=1}^{K}\left[\bar{V}_h^k(s_1^k) - \underline{V}_1^k(s_1^k)\right] \tag{37}$$

$$\leq \sum_{k=1}^{K}\sum_{h=1}^{H}\left[4\beta_h^k + \xi_h^k + \zeta_h^k\right] \tag{38}$$

For β_h^k we have

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\beta_h^k = C\sum_{h=1}^{H}\sum_{s\in S_h}\sum_{a\in A_{h,s}}\sum_{b\in B_{h,s}}\sum_{t=1}^{N_h^K(s,a,b)}\frac{1}{\sqrt{t}} \tag{39}$$

$$\leq \sqrt{KH^2|S||A||B|} \tag{40}$$

by the Cauchy-Schwarz inequality, where

$$C = \sqrt{2H^2|S|\ln(KH|S||A||B|/\delta)} \tag{41}$$

finally, this gives us

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\left[4\beta_h^k + \xi_h^k + \zeta_h^k\right] \tag{42}$$

$$\leq 4\sqrt{2KH^4|S|^2|A||B|\ln(KH|S||A||B|/\delta)} + 2\sqrt{2KH^3\ln\frac{1}{\delta}} \tag{43}$$

$$\leq 6\sqrt{2KH^4|S|^2|A||B|\ln(KH|S||A||B|/\delta)} \tag{44}$$

which completes the proof. □

Appendix B Proof of Theorem 4.2

We prove Theorem 4.2 for the max-player's exploration strategy μ^k only, as the proof for the min-player's strategy is symmetric. We first show that the upper bounds $\bar{V}_h^k$ and $\bar{Q}_h^k$ can always be achieved for some game in D_k.

Lemma B.1. *At each episode k, there exists a game $G \in D_k$ such that the upper confidence bounds $\bar{V}^k$ and $\bar{Q}^k_h$ computed by Strategic-ULCB for $\beta_t = 0$ satisfy*

$$\bar{V}^k_h(s) = \sup_\mu \inf_\nu V^{\mu,\nu}_{G,h}(s) \tag{45}$$

$$\bar{Q}^k_h(s,a,b) = \sup_\mu \inf_\nu Q^{\mu,\nu}_{G,h}(s,a,b) \tag{46}$$

for all $h \in [H]$ and $s \in S_h$, and $a \in A_{h,s}$ or $b \in B_{h,s}$.

Proof. We prove this by induction on h. Assume that for some $k \geq 1$, $h\ in[H]$, there exists a game $G \in D_k$ such that

$$\bar{V}^k_{h+1}(s) = \sup_\mu \inf_\nu V^{\mu,\nu}_{G,h+1}(s) \tag{47}$$

for all $s \in S_{h+1}$. For each $s \in S_h$, $a \in A_{h,s}$, and $b \in B_{h,s}$, if $(h,s,a,b) \in \mathcal{H}_t$, then since $G \in D^k$ we will have $\hat{R}^k_h(s,a,b) = R_{G,h}(s,a,b) = R_h(s,a,b)$ and $\hat{P}^k_h(s,a,b) = P_{G,h}(s,a,b) = P_h(s,a,b)$, and so

$$\begin{aligned}\bar{Q}^k_h(s,a,b) &= R_{G,h}(s,a,b) + P_{G,h}(s,a,b)^\top \bar{V}^k_{G,h+1} & (48)\\ &= \sup_\mu \inf_\nu Q^{\mu,\nu}_{G,h}(s,a,b) & (49)\end{aligned}$$

On the other hand, if $(h,s,a,b) \notin \mathcal{H}_t$, then we have $\bar{Q}^k_h(s,a,b) = H$. In this case, there exists a game $G' \in D^k$ that is equivalent to G for all $h' \geq h$, but for which $P_{G',h}(s,a,b,s^*) = 1$, and $R_h(s,a,b) = H$, where s^* is our hypothetical absorbing state with reward 0 for all actions and time steps. Because transition distributions can be selected independently of one another for each s, a and b, there exists $G' \in D_k$ such that $P_{G',h}(s,a,b,s^*) = 1$, and $R_h(s,a,b) = H$ for all $s \in S_h$, $a \in A_{h,s}$, and $b \in B_{h,s}$ where $(h,s,a,b) \notin \langle_t$, such that $\bar{Q}^k_h(s,a,b) = \sup_\mu \inf_\nu Q^{\mu,\nu}_{G',h}(s,a,b)$. We then have that

$$\begin{aligned}\bar{V}^k_h(s) &= \mu^k_h(s)^\top \bar{Q}^k_h(s,\cdot,\cdot)\tilde{\nu}^k_h(s) & (50)\\ &= \sup_\mu \inf_\nu \bar{Q}^k_h(s,\cdot,\cdot) & (51)\\ &= \sup_\mu \inf_\nu \bar{Q}^k_{G',h}(s,\cdot,\cdot) & (52)\end{aligned}$$

Noting that Equation 47 holds trivially for $h = H$, where we implicitly assume that $\bar{V}^k_{H+1} = V^{\mu,\nu}_{H+1} = 0$, this proves the lemma for all $h \in H$. □

To show that Strategic ULCB is strategically efficient for the max-player exploration policy, we need to show that, for some game $G \in D_k$, μ^k is the max player component of a Nash equilibrium of G.

Proof of Theorem 4.2. Let $G \in D_k$ be a game for which Equations 45 and 46 hold. By Lemma B.1, such a game always exists. We can prove that μ^k is a max-player component of an equilibrium of G by induction on h. Assume that, for $h \in [H]$ and for all $s \in S_h$

$$\mu^k \in \arg\max_{\mu} \inf_{\nu} V_{G,h+1}^{\mu,\nu}(s) \tag{53}$$

We then have that, for all $h \in H$, $s \in S_h$

$$\mu_h^k(s) \in \arg\max_{x} \inf_{y} x^\top \bar{Q}_h^k(s,\cdot,\cdot)y \tag{54}$$

$$= \arg\max_{x} \inf_{y} x^\top \left[\sup_{\mu} \inf_{\nu} Q_{G,h}^{\mu,\nu}(s,\cdot,\cdot)\right] y \tag{55}$$

$$= \arg\max_{x} x^\top \left[\sup_{\mu} \inf_{\nu} Q_{G,h}^{\mu,\nu}(s,\cdot,\cdot)\nu_h(s)\right] \tag{56}$$

$$= \arg\max_{x} x^\top \inf_{\nu} V_{G,h+1}^{\mu^k,\nu}(s) \tag{57}$$

where the last line implies that

$$\mu^k \in \arg\max_{\mu} \inf_{\nu} V_{G,h}^{\mu,\nu}(s) \tag{58}$$

Noting that Equation 53 is implicitly satisfied for $h = H$, this concludes the proof for μ^k. Repeating this process for ν^k proves the result. □

Appendix C Optimistic Nash-Q Algorithm

Algorithm 1 describes Optimistic Nash-Q algorithm of Bai et al. [2020]. Note that, in our implementation, the evaluation strategies are taken as the marginals of the most recent exploration strategy, which is itself a joint distribution over the actions for both players. Here, the sequences of learning rates α_t and exploration bonuses β_t are left as free hyperparameters that can be tuned to a specific task.

Appendix D Strategic Nash-Q Algorithm

Algorithm 2 describes the Strategic Nash-Q algorithm, which applies the strategically efficient updater rules of Strategic ULCB to the Optimistic Nash-Q algorithm of Bai et al. [2020]. Here, the sequences of learning rates α_t and exploration bonuses β_t are left as free hyperparameters that can be tuned to a specific task.

References

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

Algorithm 1 The Optimistic Nash-Q algorithm. Optimistic Nash-Q maintains upper and lower bounds on the optimal Q-function, and selects as its exploration strategy a *Coarse Correlated Equilibrium* (CCE) of the corresponding general-sum game for each state.

Inputs: $\alpha_{t\geq 0}$, $\beta_{t\geq 0}$
Initialize: $\forall(h,s,a,b)$, $\bar{Q}_h(s,a,b), \bar{V}_h(s) \leftarrow H$, $\underline{Q}_h(s,a,b), \underline{V}_h(s) \leftarrow 0$, $N_h(s,a,b) \leftarrow 0$, $\pi_h^1(s,a,b) \leftarrow 1/|A_{h,s}||B_{h,s}|$.
for episode $k = 1, \ldots, K$ **do**
  observe s_1^k.
  for step $h = 1, \ldots, H$ **do**
    take action $a_h^k \sim \mu_h^k(s_h^k)$, $b_h^k \sim \nu_h^k(s_h^k)$.
    observe reward r_h^k, next state s_{h+1}^k.
    $N_h(s_h^k, a_h^k, b_h^k) \leftarrow N_h(s_h^k, a_h^k, b_h^k) + 1$
    $t \leftarrow N_h(s_h^k, a_h^k, b_h^k)$
    $\bar{Q}_h(s_h^k, a_h^k, b_h^k) \leftarrow \min\{(1-\alpha_t)\bar{Q}_h(s_h^k, a_h^k, b_h^k) + \alpha_t(r_h^k + \bar{V}_{h+1}^k(s_{h+1}^k) + \beta_t), H\}$
    $\underline{Q}_h(s_h^k, a_h^k, b_h^k) \leftarrow \max\{(1-\alpha_t)\underline{Q}_h(s_h^k, a_h^k, b_h^k) + \alpha_t(r_h^k + \underline{V}_{h+1}^k(s_{h+1}^k) - \beta_t), 0\}$
    $\pi_h^{k+1}(s_h^k) \leftarrow \text{CCE}(\bar{Q}_h(s_h^k, \cdot, \cdot), -\underline{Q}_h(s_h^k, \cdot, \cdot))$
    $\bar{V}_h(s_h^k) \leftarrow \sum_{a\in A, b\in B} \pi_h^{k+1}(s_h^k, a, b)\bar{Q}_h(s_h^k, a, b)$
    $\underline{V}_h(s_h^k) \leftarrow \sum_{a\in A, b\in B} \pi_h^{k+1}(s_h^k, a, b)\underline{Q}_h(s_h^k, a, b)$
    $\tilde{\mu}_h^{k+1}(s_h^k) \leftarrow \sum_{b\in B} \pi_h^{k+1}(s_h^k, \cdot, b)$
    $\tilde{\nu}_h^{k+1}(s_h^k) \leftarrow \sum_{a\in A} \pi_h^{k+1}(s_h^k, a, \cdot)$
  end for
end for

Yu Bai, Chi Jin, and Tiancheng Yu. Near-optimal reinforcement learning with self-play. *Advances in Neural Information Processing Systems*, 33, 2020.

Tsachy Weissman, Erik Ordentlich, Gadiel Seroussi, Sergio Verdu, and Marcelo J Weinberger. Inequalities for the l1 deviation of the empirical distribution. *Hewlett-Packard Labs, Tech. Rep*, 2003.

Algorithm 2 The Strategic Nash-Q algorithm. Unlike Optimistic Nash-Q, Strategic Nash-Q computes the max and min-player policies for each state independently, and updates its value function bounds under the assumption that the adversary acts pessimistically (optimizing the lower-bound on its expected return, rather than the upper bound). Like Strategic ULCB, Strategic Nash-Q maintains separate evaluation policies $\tilde{\mu}^k$ and $\tilde{\nu}^k$.

Inputs: $\alpha_{t\geq 0}$, $\beta_{t\geq 0}$
Initialize: $\forall(h,s,a,b)$, $\bar{Q}_h(s,a,b) \leftarrow H$, $\underline{Q}_h(s,a,b) \leftarrow 0$, $N_h(s,a,b) \leftarrow 0$, $\mu^1_h(s,a) \leftarrow {}^1/_{|A_{h,s}|}$, $\nu^1_h(s,a) \leftarrow {}^1/_{|B_{h,s}|}$.
for episode $k = 1,\ldots,K$ **do**
  observe s^k_1.
  for step $h = 1,\ldots,H$ **do**
    take action $a^k_h \sim \mu^k_h(s^k_h)$, $b^k_h \sim \nu^k_h(s^k_h)$.
    observe reward r^k_h, next state s^k_{h+1}.
    $N_h(s^k_h,a^k_h,b^k_h) \leftarrow N_h(s^k_h,a^k_h,b^k_h) + 1$
    $t \leftarrow N_h(s^k_h,a^k_h,b^k_h)$
    $\bar{Q}_h(s^k_h,a^k_h,b^k_h) \leftarrow \min\{(1-\alpha_t)\bar{Q}_h(s^k_h,a^k_h,b^k_h)+\alpha_t(r^k_h+\bar{V}^k_{h+1}(s^k_{h+1})+\beta_t), H\}$
    $\underline{Q}_h(s^k_h,a^k_h,b^k_h) \leftarrow \max\{(1-\alpha_t)\underline{Q}_h(s^k_h,a^k_h,b^k_h)+\alpha_t(r^k_h+\underline{V}^k_{h+1}(s^k_{h+1})-\beta_t), 0\}$
    $\mu^{k+1}_h(s^k_h), \tilde{\nu}^{k+1}_h(s^k_h) \leftarrow \mathrm{Nash}(\bar{Q}_h(s^k_h,\cdot,\cdot), -\bar{Q}_h(s^k_h,\cdot,\cdot))$
    $\tilde{\mu}^{k+1}_h(s^k_h), \nu^{k+1}_h(s^k_h) \leftarrow \mathrm{Nash}(\underline{Q}_h(s^k_h,\cdot,\cdot), -\underline{Q}_h(s^k_h,\cdot,\cdot))$
    $\bar{V}^k_h(s) \leftarrow \mu^k_h(s)^\top \bar{Q}^k_h(s,\cdot,\cdot)\tilde{\nu}^k_h$
    $\underline{V}^k_h(s) \leftarrow \tilde{\mu}^k_h(s)^\top \underline{Q}^k_h(s,\cdot,\cdot)\nu^k_h$
  end for
end for