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 \( \beta_t \) as

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

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 \( V \in [0,H]|S| \)

\[
|\hat{P}_h^k(s,a,b)^\top V - P(s,a,b)^\top V| \leq \beta_t
\]

for \( t = N_k^k(s,a,b) \).

Proof. When \( N_k^k(s,a,b) = 0 \), 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

\[
\Pr \left\{ \| \hat{P}_h^k(s,a,b) - P(s,a,b) \|_1 \geq \epsilon \right\} \leq (2|S| - 2) \exp\left\{ -N_h^k(s,a,b)\epsilon^2 \right\}
\]

Note that, for all \( V \in [0,H]|S| \)

\[
|\hat{P}_h^k(s,a,b)^\top V - P(s,a,b)^\top V| \leq H\|\hat{P}_h^k(s,a,b) - P(s,a,b)\|_1
\]

and so for \( t = N_k^k(s,a,b) \) and \( \beta_t \) defined according to Equation 1 we have

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

Taking the union bound over \( k, h, s, a \) and \( b \) yields the desired result. \( \square \)

For games with deterministic transitions, \( \hat{P}_h^k(s,a,b) = P(s,a,b) \) whenever \( N_k^k(s,a,b) > 0 \), 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 \( \hat{V}_h^k \) and \( \hat{V}_k^k \) 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 $\beta_t$ as defined in Equation 1, the for all $k \in [K]$, $h \in H$ and $s \in S_h$, we have

$$V_h^k(s) \geq \sup_{\mu} V_h^{\mu,\nu}(s) \quad (6)$$
$$V_h^k(s) \leq \inf_{\nu} V_h^{\mu,\nu}(s) \quad (7)$$

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 $s \in S_h$

$$V_{h+1}^k(s) \geq \sup_{\mu} V_{h+1}^{\mu,\nu}(s) \quad (8)$$

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

$$Q^k_h(s,a,b) = R(s,a,b) \geq P(s,a,b) V_{h+1}^k$$

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

$$V_{h+1}^k(s) - \sup_{\mu} V_{h+1}^{\mu,\nu}(s) = \mu_h^k(s) V_h^k(s,a,b) \nu_h^k(s)$$

which proves the inductive step. The first inequality follows directly from Equation 9, while the second inequality follows from the fact that $(\mu_h^k(s), \nu_h^k(s))$ for a Nash equilibrium of the matrix game defined by $Q_h^k(s,a,b)$, and so $\mu_h^k(s)$ is a best-response to $\nu_h^k(s)$ under $Q_h^k(s,a,b)$. Finally, we can see that Equation 8 holds trivially for $h = H + 1$, where we implicitly assume that $V_h^k(s) = \sup_\nu V_h^{\mu,\nu}(s) = 0$, which concludes the proof. □
Lemma A.2 will be sufficient to prove Theorem 4.1 and bound the NashConv regret of the evaluation policies \( \hat{\mu}^k \) and \( \hat{\nu}^k \). 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

\[
\text{Regret}(K) = \sum_{k=1}^{K} \sup_{\mu} V_{1}^{\mu,k} (s_1) - \inf_{\nu} V_{1}^{\mu,k,\nu} (s_1)
\]  

(18)

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

\[
\sup_{\mu} V_{1}^{\mu,k} (s_h^k) - \inf_{\nu} V_{1}^{\mu,k,\nu} (s_h^k)
\]

(19)

\[
\leq \hat{V}_h^k (s_h^k) - V_h^k (s_h^k)
\]

(20)

\[
= \mu_h^k(s_h^k)^{\top} Q_h^k(s_h^k, \cdot, \cdot) \nu_h^k(s_h^k) - \hat{\mu}_h^k(s_h^k)^{\top} Q_h^k(s_h^k, \cdot, \cdot) \nu_h^k(s_h^k)
\]

(21)

\[
\leq \mu_h^k(s_h^k)^{\top} \left[ \hat{Q}_h^k(s_h^k, \cdot, \cdot) - Q_h^k(s_h^k, \cdot, \cdot) \right] \nu_h^k(s_h^k)
\]

(22)

\[
= \mu_h^k(s_h^k)^{\top} \left[ \hat{Q}_h^k(s_h^k, \cdot, \cdot) - Q_h^k(s_h^k, \cdot, \cdot) \right] \nu_h^k(s_h^k) + 2\beta_h + \xi_h
\]

(23)

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

\[
\mu_h^k(s_h^k)^{\top} \left[ \hat{Q}_h^k(s_h^k, \cdot, \cdot) - Q_h^k(s_h^k, \cdot, \cdot) \right] \nu_h^k(s_h^k)
\]

(24)

\[
= \left[ \hat{Q}_h^k - Q_h^k \right] (s_h^k, a_h^k, b_h^k) + \xi_h
\]

(25)

\[
= \left[ \hat{V}_h^k - V_h^k \right] + 2\beta_h + \xi_h
\]

(26)

\[
= P(s_h^k, a_h^k, b_h^k)^{\top} \left[ \hat{V}_h^k - V_h^k \right] + 4\beta_h + \xi_h
\]

(27)

\[
= \left[ \hat{V}_{h+1}^k - V_{h+1}^k \right] (s_{h+1}^k) + 4\beta_h + \xi_h
\]

(28)

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

\[
\xi_h^k = E_{a, b \sim \mu_h^k(s_h^k)} \left[ Q_h^k - \hat{Q}_h^k \right] (s_h^k, a, b)
\]

(29)

\[
- \left[ \hat{Q}_h^k - Q_h^k \right] (s_h^k, a_h^k, b_h^k)
\]

(30)

\[
\zeta_h^k = E_{s \sim P(s_h^k, a_h^k, b_h^k)} \left[ \hat{V}_h^k - V_h^k \right] (s)
\]

(31)

\[
- \left[ \hat{V}_{h+1}^k - V_{h+1}^k \right] (s_{h+1}^k)
\]

(32)

Here \( \xi_h^k \) and \( \zeta_h^k \) are not i.i.d., but the sequences of their partial sums over \( k \) and
are martingales, and so by the Azuma-Hoeffding inequality

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

we then have

\[
\sum_{k=1}^{K} \sup_{\mu} V_{1,\mu}^k (s_1) - \inf_{\nu} V_{1,\nu}^k (s_1) \leq \sum_{k=1}^{K} [\tilde{V}_h^k (s_1) - V_1^k (s_1)] \tag{36}
\]

For \( \beta_h^k \) we have

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

by the Cauchy-Schwarz inequality, where

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

finally, this gives us

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

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

which completes the proof.

**Appendix B  Proof of Theorem 4.2**

We prove Theorem 4.2 for the max-player’s exploration strategy \( \mu^k \) only, as the proof for the min-player’s strategy is symmetric. We first show that the upper bounds \( \tilde{V}_h^k \) and \( \tilde{Q}_h^k \) 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 $\hat{V}^k_h$ and $\hat{Q}^k_h$ computed by Strategic-ULCB for $\beta_t = 0$ satisfy

$$\hat{V}^k_h(s) = \sup_\mu \inf_\nu V^\mu_\nu_G(s)$$ \hfill (45)

$$\hat{Q}^k_h(s, a, b) = \sup_\mu \inf_\nu Q^\mu_\nu_G(s, a, b)$$ \hfill (46)

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

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

$$\hat{V}^k_{h+1}(s) = \sup_\mu \inf_\nu V^\mu_\nu_{G_{h+1}}(s)$$ \hfill (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 $(h, s, a, b) \in H_t$, then since $G \in D_k$ we will have $R^k_h(s, a, b) = R_G(s, a, b) = R_h(s, a, b)$ and $P^k_h(s, a, b) = P_{G,h}(s, a, b) = P_h(s, a, b)$, and so

$$\hat{Q}^k_h(s, a, b) = R_{G,h}(s, a, b) + P_{G,h}(s, a, b) \hat{V}^k_{G,h+1}$$ \hfill (48)

$$\sup_\mu \inf_\nu Q^\mu_\nu_{G,h}(s, a, b)$$ \hfill (49)

On the other hand, if $(h, s, a, b) \notin H_t$, then we have $\hat{Q}^k_h(s, a, b) = H$. 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 $(h, s, a, b) \notin H_t$, such that $\hat{Q}^k_h(s, a, b) = \sup_\mu \inf_\nu Q^\mu_\nu_{G'}(s, a, b)$. We then have that

$$\hat{V}^k_h(s) = \mu^k_h(s) \hat{Q}^k_h(s, \cdot, \cdot) \hat{P}^k_h(s)$$ \hfill (50)

$$\sup_\mu \inf_\nu \hat{Q}^k_h(s, \cdot, \cdot)$$ \hfill (51)

$$\sup_\mu \inf_\nu \hat{Q}^k_{G'}(s, \cdot, \cdot)$$ \hfill (52)

Noting that Equation 47 holds trivially for $h = H$, where we implicitly assume that $\hat{V}^k_{H+1} = V^\mu_\nu_{H+1} = 0$, this proves the lemma for all $h \in H$. \hfill \qed

To show that Strategic ULCB is strategically efficient for the max-player exploration policy, we need to show that, for some game $G \in D_k$, $\mu^k$ 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 $\mu^k$ is a
max-player component of an equilibrium of $G$ by induction on $h$. Assume that, for $h \in [H]$ and for all $s \in S_h$

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

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

$$\mu^k_h(s) \in \arg\max_{x} \inf_{y} x^\top \hat{Q}^{k}_h(s, \cdot, \cdot) y \quad (54)$$

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

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

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

where the last line implies that

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

Noting that Equation 53 is implicitly satisfied for $h = H$, this concludes the proof for $\mu^k$. Repeating this process for $\nu^k$ proves the result. \qed

### 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 $\alpha_t$ and exploration bonuses $\beta_t$ 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 $\alpha_t$ and exploration bonuses $\beta_t$ are left as free hyperparameters that can be tuned to a specific task.

### References

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: $\alpha \geq 0$, $\beta \geq 0$
Initialize: $\forall (h, s, a, b), \bar{Q}_h(s, a, b), \bar{V}_h(s) \leftarrow H$, $Q_h(s, a, b), V_h(s) \leftarrow 0$, $N_h(s, a, b) \leftarrow 0$, $\pi^1_h(s, a, b) \leftarrow 1/|A_h| |B_h|$. 

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 $s^k_{h+1}$.

$t \leftarrow N_h(s^k_h, a^k_h, b^k_h)$

$Q_h(s^k_h, a^k_h, b^k_h) \leftarrow \min \{(1-\alpha_t)Q_h(s^k_h, a^k_h, b^k_h) + \alpha_t(r^k_h + V^k_{h+1}(s^k_{h+1}) + \beta_t), H\}$

$Q_h(s^k_h, a^k_h, b^k_h) \leftarrow \max \{(1-\alpha_t)Q_h(s^k_h, a^k_h, b^k_h) + \alpha_t(r^k_h + V^k_{h+1}(s^k_{h+1}) - \beta_t), 0\}$

$\pi^{k+1}_h(s^k_h) \leftarrow \text{CCE}(\bar{Q}_h(s^k_h, \cdot, \cdot), -\bar{Q}_h(s^k_h, \cdot, \cdot))$

$\bar{V}_h(s^k_h) \leftarrow \sum_{a \in A, b \in B} \pi^{k+1}_h(s^k_h, a, b) \bar{Q}_h(s^k_h, a, b)$

$V_h(s^k_h) \leftarrow \sum_{a \in A, b \in B} \pi^{k+1}_h(s^k_h, a, b) Q_h(s^k_h, a, b)$

$\mu^{k+1}_h(s^k_h) \leftarrow \sum_{b \in B} \pi^{k+1}_h(s^k_h, \cdot, b) \bar{Q}_h(s^k_h, a, b)$

$\nu^{k+1}_h(s^k_h) \leftarrow \sum_{a \in A} \pi^{k+1}_h(s^k_h, a, \cdot)$

end for

end for


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:** \( \forall (h, s, a, b) \), $Q_h(s, a, b) \leftarrow H$, $Q_h(s, a, b) \leftarrow 0$, $N_h(s, a, b) \leftarrow 0$, $\mu_1^h(s, a) \leftarrow 1/\vert A_{h,s} \vert$, $\nu_1^h(s, a) \leftarrow 1/\vert B_{h,s} \vert$.

**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 $s^{k+1}_h$.

$N_h(s^k_h, a^k_h, b^k_h) \leftarrow N_h(s^k_h, a^k_h, b^k_h) + 1$

$t \leftarrow N_h(s^k_h, a^k_h, b^k_h)$

$Q_h(s^k_h, a^k_h, b^k_h) \leftarrow \min \{(1-\alpha_t)Q_h(s^k_h, a^k_h, b^k_h) + \alpha_t(r^k_h + \bar{V}^k_{h+1}(s^{k+1}_h) + \beta_t), H\}$

$Q_h(s^k_h, a^k_h, b^k_h) \leftarrow \max \{(1-\alpha_t)Q_h(s^k_h, a^k_h, b^k_h) + \alpha_t(r^k_h + \bar{V}^k_{h+1}(s^{k+1}_h) - \beta_t), 0\}$

$\mu^{k+1}_h(s^k_h, \cdot, \cdot, \cdot) \leftarrow \text{Nash}(Q_h(s^k_h, \cdot, \cdot, \cdot), -Q_h(s^k_h, \cdot, \cdot, \cdot))$

$\nu^{k+1}_h(s^k_h, \cdot, \cdot, \cdot) \leftarrow \text{Nash}(Q_h(s^k_h, \cdot, \cdot, \cdot), -Q_h(s^k_h, \cdot, \cdot, \cdot))$

$\bar{V}^k_h(s) \leftarrow \mu^{k}_h(s) \top Q^k_h(s, \cdot, \cdot)\tilde{\nu}^k_h$

$\bar{V}^k_h(s) \leftarrow \mu^{k}_h(s) \top Q^k_h(s, \cdot, \cdot)\tilde{\nu}^k_h$

**end for**

**end for**