1 A Motivating Example: Climb Game

We analyze a fully-cooperative matrix game known as Climb Game. In Section 1.1, we show how popular exploration strategies, including unstructured strategies like uniform exploration and task-specific strategies like $\epsilon-$greedy, fail to efficiently explore the climb game. By contrast, we show in Section 1.2 that a simple structured exploration strategy can substantially improve the exploration efficiency.

A climb game $G_f(n, u, U)$ is a n-player game where $\mathcal{A}_i = \{0, \dots, U-1\}$ for any player i. The reward of a joint action $\boldsymbol{a} \in \mathcal{A}$ is determined by the number of players performing a specific action u (denoted as $\#u$), which is

$$R(\boldsymbol{a}) = \begin{cases} 1, & \text{if } \#u = n, \\ 1 - \delta \ (0 < \delta < 1), & \text{if } \#u = 0, \ . \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

1.1 Exploration Challenge

A climb game $G_f(n, u, U)$ has three groups of NE: the Pareto optimal NE $(u, u, \dots, u)$, the sub-optimal NEs $\{(a_1, a_2, \dots, a_n) \mid \forall i, \ a_i \neq u\}$, and the zero-reward NEs $\{(a_1, a_2, \dots, a_n) \mid 1 < \#u < n\}$. The sheer difference in size of the three subsets of NEs makes it particularly challenging for RL agents to learn the optimal policy profile without sufficient exploration, as evidenced by the theoretical analysis below and empirical evaluation in experiment Section.

Consider a 2-agent climb game $G_f(2, 0, U)$. A joint action $\boldsymbol{a}$ can be represented by a pair of one-hot vectors $[\mathbf{e}_i, \mathbf{e}_j] \in \{0, 1\}^{2U}$. Let $q(\mathbf{x}, \mathbf{y}; \theta)$ be a joint Q function parameterized by θ that takes input $\mathbf{x}, \mathbf{y} \in \{0, 1\}^U$ and is learned to approximate the reward of the game. We hope the joint Q function has the same optimal policy profile.

Definition 1.1. *We call a joint Q function* $q(\mathbf{x}, \mathbf{y}; \theta)$ equivalently optimal *when* $q(\mathbf{e}_0, \mathbf{e}_0; \theta) = \max_{0 \leq i,j < U} q(\mathbf{e}_i, \mathbf{e}_j; \theta)$. *When a joint Q function is equivalently optimal, we think it finds the correct optimal policy.*

Since neural networks are difficult to analyze in general [**?**], we parameterize the joint Q function in a quadratic form:

$$q(\mathbf{x}, \mathbf{y}; \mathbf{W}, \mathbf{b}, \mathbf{c}, d) = \mathbf{x}^\top \mathbf{W} \mathbf{y} + \mathbf{b}^\top \mathbf{x} + \mathbf{c}^\top \mathbf{y} + d \tag{2}$$

A Gaussian prior $p(\mathbf{W}) = \mathcal{N}(\mathbf{W}; 0, \sigma_w^2 I)$ is introduced under the assumption that a non-linear $\mathbf{W}$ is harder and slower to learn. Quadratic functions have been used in RL [**?**, **?**] as a replacement of commonly-used multi-layer perceptron, and there are also theoretical results [**?**] analyzing neural networks with quadratic activation. For the climb game, it is easy to verify that the quadratic coefficients make the joint Q function sufficiently expressive to perfectly fit the reward function. Therefore, the learning process of Q is mainly affected by how the exploration policy samples the data.

Consider an exploration policy $p_e^{(t)}$ that selects joint action $\boldsymbol{a} = (i, j)$ at step t with probability $p_e^{(t)}(i, j)$. The efficiency of an exploration policy can be measured by the required number of steps for learning an equivalently optimal Q function using the maximum likelihood estimator over the data sampled from $p_e^{(t)}$. The learning objective includes both the prior $p(\mathbf{W})$ and the likelihood of prediction error $E_{ij} = q(\mathbf{e}_i, \mathbf{e}_j; \cdot) - R_{ij}$. Here the prediction error is depicted by a Gaussian distribution $p(E_{ij}) = \mathcal{N}(E_{ij}; 0, \sigma_e^2)$ for every visited joint action (i, j). We use Gaussian distribution because maximizing the log-likelihood is equivalent to minimizing the square of prediction error. Therefore, the learning objective for the Q function can be formulated as:

$$
\begin{aligned}
&\mathcal{J}^{(T)}(\mathbf{W}, \mathbf{b}, \mathbf{c}, d) \\
=&\mathbb{E}_{\{(i^{(t)}, j^{(t)}) \sim p_e^{(t)}\}_{t=1}^{T}} \log \left(p(\mathbf{W}) \prod_{t'=1}^{T} p(E_{i^{(t)} j^{(t)}}) \right) \\
=&\sum_{t=1}^{T} \mathbb{E}_{(i,j) \sim p_e^{(t)}} \left[\log \mathcal{N}(q(\mathbf{e}_i, \mathbf{e}_j; \mathbf{W}, \mathbf{b}, \mathbf{c}, d) - R_{ij}; 0, \sigma_e^2) \right] \\
&+ \log \mathcal{N}(\mathbf{W}; 0, \sigma_w^2 I) + \text{Const.}
\end{aligned} \tag{3}
$$

We use $q_{\mathcal{J}^{(T)}}(\mathbf{W}, \mathbf{b}, \mathbf{c}, d)$ to denote the learned joint Q function that maximizes $\mathcal{J}^{(T)}$ at step T. $q_{\mathcal{J}^{(T)}}(\mathbf{W}, \mathbf{b}, \mathbf{c}, d)$ is determined by the exploration policy $p_e^{(t)}$ and the exploration steps T. Then we have the following theorem for the uniform exploration strategy.

Theorem 1.2 (uniform exploration)**.** *Assume $\delta \leq \frac{1}{6}, U \geq 3$. Using a uniform exploration policy in the climb game $G_f(2, 0, U)$, $q_{\mathcal{J}^{(T)}}(\mathbf{W}, \mathbf{b}, \mathbf{c}, d)$ will become equivalently optimal only after $T = \Omega(|\mathcal{A}|\delta^{-1})$ steps. When $\delta = 1$, $T = O(1)$ steps suffice to learn the equivalently optimal joint Q function, suggesting the inefficiency of uniform exploration is due to a large set of sub-optimal NEs.*

The intuition behind Theorem 1.2 is that the hardness of exploration in climb games largely comes from the sparsity of solutions: a set of sub-optimal NEs exist but there is only a single Pareto optimal NE. Learning the joint Q function can be influenced by the sub-optimal NEs. And if the exploration attempts are not well coordinated, a lot of zero reward would be encountered, making it hard to find the Pareto optimal NE. We also remark that uniform exploration can be particularly inefficient since the term $|\mathcal{A}|$ can be exponentially large in a multi-agent system. So we may need to possibly reduce the search space and identify a smaller "critical" subspace for more efficient exploration.

Next, we consider the case of another popular exploration paradigm, ϵ-greedy exploration.

Theorem 1.3 (ϵ-greedy exploration)**.** *Assume $\delta \leq \frac{1}{32}, U \geq \max(4, \sigma_w \sigma_e^{-1})$. In the climb game $G_f(2, 0, U)$, under ϵ-greedy exploration with fixed $\epsilon \leq \frac{1}{2}$, $q_{\mathcal{J}^{(T)}}(\mathbf{W}, \mathbf{b}, \mathbf{c}, d)$ will become equivalently optimal only after $T = \Omega(|\mathcal{A}|\delta^{-1}\epsilon^{-1})$ steps. If $\epsilon(t) = 1/t$, it requires $T = \exp\left(\Omega\left(|\mathcal{A}|\delta^{-1}\right)\right)$ exploration steps to be equivalently optimal.*

By comparing 1.2 and 1.3, ϵ-greedy results in even poorer exploration efficiency than uniform exploration. Note the ϵ-greedy strategy is training policy specific, i.e., the exploration behavior varies as the training policy changes. Theorem 1.3 suggests that when the policy is sub-optimal, the induced ϵ-greedy exploration strategy can be even worse than uniform exploration. Hence, it can be beneficial to adopt a separate exploration independent from the training policy.

The above analysis shows that common exploration strategies like uniform exploration or ϵ-greedy exploration are inefficient. The difficulty of exploration in the climb game is that it requires coordination between different agents to reach high-rewarding states, but naive exploration strategies lack cooperation between agents.

1.2 Structured Exploration

We will show that it is possible to design a better exploration strategy with some prior knowledge of the climb game structure. Consider a specific structured exploration strategy $p_e^{(t)}(i,j) = U^{-1}[\mathbb{1}_{i=j}]$, where both agents always choose the same action. With such a strategy, we can quickly find the optimal solution to the game. More formally, we have the following theorem.

Theorem 1.4 (structured exploration)**.** *In the climb game $G_f(2,0,U)$, under structured exploration $p_e^{(t)}(i,j) = U^{-1}[\mathbb{1}_{i=j}]$, $q_{\mathcal{J}^{(T)}}(\mathbf{W},\mathbf{b},\mathbf{c},d)$ is equivalently optimal at step $T = O(1)$.*

Theorem 1.4 shows the efficiency of exploration can be greatly improved if the exploration strategy captures a proper structure of the problem, i.e., all agents taking the same action. We further remark that by considering a set of similar climb games $\mathcal{G} = \{G_f(2,u,U)\}_{u=0}^{U-1}$, the structured exploration strategy $p_e^{(t)}(i,j) = U^{-1}[\mathbb{1}_{i=j}]$ can be interpreted as a uniform distribution over the optimal policies of this game set $\mathcal{G}$. This interesting fact suggests that we can first collect a set of similarly structured games and then derive effective exploration strategies from these similar games. Once a set of structured exploration strategies are collected, we can further adopt them for fast learning in a novel game with a similar problem structure. We take the inspiration here and develop a general meta-exploration algorithm in the next section.

2 Proof

2.1 Proof of Lemma on Equivalent Optimality

Lemma 2.1. *In the 2-agent Climb Game with single-agent action space $|\mathcal{A}| = U$ and reward matrix*

$$R = \begin{pmatrix} r & 0 & \cdots & 0 \\ 0 & r(1-\delta) & \cdots & r(1-\delta) \\ \vdots & \vdots & \ddots & \vdots \\ 0 & r(1-\delta) & \cdots & r(1-\delta) \end{pmatrix},$$

for any exploration policy p_e where $p_e^{(t)}(i,j)$ is the probability of trying action (i,j) at time step t, given the objective function

$$\begin{aligned} &\mathcal{J}^{(T)}(\mathbf{W}, \mathbf{b}, \mathbf{c}, d) \\ =&\sum_{t=1}^{T} \mathbb{E}_{(i,j)\sim p_e^{(t)}} \left[\log \mathcal{N}(q(\mathbf{e}_i, \mathbf{e}_j; \mathbf{W}, \mathbf{b}, \mathbf{c}, d) - R_{ij}; 0, \sigma_e^2)\right] \\ &+ \log \mathcal{N}(W; 0, \sigma_W^2 I) + \textit{Constant} \end{aligned} \tag{4}$$

maximized by parameters $\mathbf{W}^*, \mathbf{b}^*, \mathbf{c}^*, d^*$, the joint Q function $q(\mathbf{e}_i, \mathbf{e}_j; \mathbf{W}^*, \mathbf{b}^*, \mathbf{c}^*, d^*)$ is equivalently optimal if the following criterion holds

$$r\delta \geq \left(\frac{f_2}{f_0} - 1\right) \frac{m^2}{f_2\lambda + m^2} \frac{r(2-\delta)}{1 + \frac{2f_2(f_1\lambda+2m)m}{f_1(f_2\lambda+m^2)} + \frac{f_2(f_0\lambda+1)m^2}{f_0(f_2\lambda+m^2)}}. \tag{5}$$

Here we use

$$\begin{aligned} f_0 =& \frac{1}{T}\sum_{t=1}^{T} p_e^{(t)}(0,0) \\ f_1 =& \frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{U-1} \left(p_e^{(t)}(0,i) + p_e^{(t)}(i,0)\right) \\ f_2 =& \frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{U-1}\sum_{j=1}^{U-1} p_e^{(t)}(i,j) \\ m =& U-1 \\ \lambda =& \frac{T\sigma_w^2}{\sigma_e^2} \end{aligned}$$

for a clearer demonstration of the criterion.

Proof. From the symmetry of the parameters and the concavity of the objective

function, $\exists W_0, W_1, W_2, B, C, D$ such that

$$\begin{aligned}
W_0 =& \mathbf{W}^*_{00} \\
W_1 =& \mathbf{W}^*_{0i} = \mathbf{W}^*_{i0}, && \forall i \neq 0 \\
W_2 =& \mathbf{W}^*_{ij}, && \forall i, j \neq 0 \\
B =& \mathbf{b}^*_0 = \mathbf{c}^*_0 \\
C =& \mathbf{b}^*_i = \mathbf{c}^*_i, && \forall i \neq 0 \\
D =& d
\end{aligned}$$

Rewrite the objective function (4) we obtain

$$\begin{aligned}
\mathcal{J} = & -\frac{T}{2\sigma_e^2}\left(f_0(W_0 + 2B + D - r)^2\right. \\
& + f_1(W_1 + B + C + D)^2 \\
& \left. + f_2(W_2 + 2C + D - r(1-\delta))^2\right) \\
& - \frac{1}{2\sigma_w^2}\left(W_0^2 + 2mW_1^2 + m^2 W_2^2\right)
\end{aligned} \tag{6}$$

Further, let

$$\begin{aligned}
K_0 =& 2B + D - r \\
K_1 =& B + C + D \\
K_2 =& 2C + D - r(1-\delta)
\end{aligned}$$

and immediately

$$K_0 + K_2 = 2K_1 - r(2-\delta) \tag{7}$$

Following equation (6),

$$\begin{aligned}
\left(2\frac{\partial}{\partial W_0} + \frac{\partial}{\partial W_1} - \frac{\partial}{\partial B}\right)\mathcal{J} = 0 \Rightarrow & W_0 = -mW_1 \\
\left(2\frac{\partial}{\partial W_2} + \frac{\partial}{\partial W_1} - \frac{\partial}{\partial C}\right)\mathcal{J} = 0 \Rightarrow & W_1 = -mW_2 \\
\frac{\partial}{\partial W_0}\mathcal{J} = 0 \Rightarrow & W_0 = -\frac{f_0\lambda}{f_0\lambda + 1}K_0 \\
\frac{\partial}{\partial W_1}\mathcal{J} = 0 \Rightarrow & W_1 = -\frac{f_1\lambda}{f_1\lambda + 2m}K_1 \\
\frac{\partial}{\partial W_2}\mathcal{J} = 0 \Rightarrow & W_2 = -\frac{f_2\lambda}{f_2\lambda + m^2}K_2 \\
\frac{\partial}{\partial B}\mathcal{J} = 0 \Rightarrow & \frac{W_0 + K_0}{W_1 + K_1} = -\frac{f_1}{2f_0} \\
\frac{\partial}{\partial C}\mathcal{J} = 0 \Rightarrow & \frac{W_1 + K_1}{W_2 + K_2} = -\frac{2f_2}{f_1}
\end{aligned}$$

and together with equation (7), we obtain

$$K_2 = \frac{r(2-\delta)}{1 + \frac{2f_2(f_1\lambda+2m)m}{f_1(f_2\lambda+m^2)} + \frac{f_2(f_0\lambda+1)m^2}{f_0(f_2\lambda+m^2)}}. \tag{8}$$

Finally we deduce the criterion

$$\begin{aligned}
&W_0 + 2B + D \geq W_2 + 2C + D \\
\Leftrightarrow &r\delta \geq \left(1 - \frac{f_2}{f_0}\right)(W_2 + K_2) \\
\Leftrightarrow &r\delta \geq \left(\frac{f_2}{f_0} - 1\right) \frac{m^2}{f_2\lambda + m^2} \frac{r(2-\delta)}{1 + \frac{2f_2(f_1\lambda+2m)m}{f_1(f_2\lambda+m^2)} + \frac{f_2(f_0\lambda+1)m^2}{f_0(f_2\lambda+m^2)}}.
\end{aligned}$$

□

2.2 Proof for Theorem 1.2 (uniform exploration)

Theorem 1.2. *Assume $\delta \leq \frac{1}{6}, U \geq 3$. In the Climb Game $G_f(2, 0, U)$, given the quadratic joint Q function form $q(\mathbf{x}, \mathbf{y}; \mathbf{W}, \mathbf{b}, \mathbf{c}, d)$ and a Gaussian prior $p(\mathbf{W}) = \mathcal{N}(\mathbf{W}; 0, \sigma_w^2 I)$, using a uniform exploration policy, $q_{\mathcal{J}^{(T)}}(\mathbf{W}, \mathbf{b}, \mathbf{c}, d)$ will become equivalently optimal only after $T = \Omega(|\mathcal{A}|\delta^{-1})$ steps. When $\delta = 1$, $T = O(1)$ steps suffice to learn the equivalently optimal joint Q function, meaning the inefficiency of uniform exploration is due to a large set of suboptimal NEs.*

Proof. Under uniform exploration,

$$f_0 = \frac{1}{U^2}, f_1 = \frac{2m}{U^2}, f_2 = \frac{m^2}{U^2}.$$

Criterion (5) can be reformulated to

$$\delta \geq \frac{(m^2-1)(2-\delta)}{\left(1 + \frac{\lambda}{U^2}\right)(m+1)^2} \tag{9}$$

and thus with $\lambda = \frac{T\sigma_w^2}{\sigma_e^2}, m \geq 2, \delta \leq \frac{1}{6}$,

$$\begin{aligned}
T \geq &\frac{U^2\sigma_e^2}{\sigma_w^2}\left(\frac{(m^2-1)(2-\delta)}{(m+1)^2} - 1\right) \\
\geq &\frac{U^2\sigma_e^2}{\sigma_w^2}\left(\frac{3}{\delta} - \frac{6}{\delta}\right) \\
= &\frac{U^2\sigma_e^2}{6\sigma_w^2\delta}
\end{aligned}$$

On the other hand, in non-penalty Climb Game where $\delta = 1$, if at any time step $\exists(i,j) \neq (0,0)$ where the joint Q function $q_{\mathcal{J}^{(T)}}$ weighs action (a_i, a_j) more than the action (a_0, a_0), just swap the parameters related to (i,j) with thos related to $(0,0)$ and the objective function $\mathcal{J}^{(T)}$ will be increased, which makes a contradiction. Hence, $T = 1$ suffices for the non-penalty Climb Game.

□

2.3 Proof for Theorem 1.3 (ϵ-greedy exploration)

Theorem 1.3 *Assume $\delta \leq \frac{1}{32}, U \geq \max(4, \sigma_w \sigma_e^{-1})$. In the Climb Game $G_f(2, 0, U)$, given the quadratic joint Q function form $q(\mathbf{x}, \mathbf{y}; \mathbf{W}, \mathbf{b}, \mathbf{c}, d)$ and a Gaussian prior $p(\mathbf{W}) = \mathcal{N}(\mathbf{W}; 0, \sigma_w^2 I)$, under ϵ-greedy exploration with fixed $\epsilon \leq \frac{1}{2}$, $q_{\mathcal{J}^{(T)}}(\mathbf{W}, \mathbf{b}, \mathbf{c}, d)$ will become equivalently optimal only after $T = \Omega(|\mathcal{A}|\delta^{-1}\epsilon^{-1})$ steps. If $\epsilon(t) = 1/t$, it requires $T = \exp\left(\Omega\left(|\mathcal{A}|\delta^{-1}\right)\right)$ exploration steps to be equivalently optimal.*

Proof. Under the circumstances here, after the first step of uniform exploration, the sub-optimal policy will be used for ϵ-greedy exploration. Then for both fixed ϵ or linearly decaying ϵ, the following always holds:

$$\begin{aligned}
\frac{f_1}{f_0} =& 2m \\
\frac{f_2}{f_0} \geq& \max(2, m^2) \\
f_2 \geq& \min(1-\epsilon, m^2 U^{-2}) \geq \frac{1}{2}.
\end{aligned}$$

Then it can be derived from criterion 5 that

$$\begin{aligned}
r\delta \geq& \left(\frac{f_2}{f_0} - 1\right) \frac{m^2}{f_2\lambda + m^2} \frac{r(2-\delta)}{1 + \frac{2f_2(f_1\lambda+2m)m}{f_1(f_2\lambda+m^2)} + \frac{f_2(f_0\lambda+1)m^2}{f_0(f_2\lambda+m^2)}} \\
=& \left(\frac{f_2}{f_0} - 1\right) \frac{m^2}{f_2\lambda + m^2} \frac{r(2-\delta)}{1 + \frac{f_2(f_0\lambda+1)(m^2+2m)}{f_0(f_2\lambda+m^2)}} \\
=& \left(\frac{f_2}{f_0} - 1\right) \frac{m^2}{f_2\lambda + m^2} \frac{r(2-\delta)}{(m+1)^2} \\
\geq& (m^2-1) \frac{m^2}{\lambda + m^2} \frac{r(2-\delta)}{(m+1)^2}
\end{aligned} \tag{10}$$

Similar to inequality (9), this yields to

$$\lambda \geq \frac{m^2}{6\delta} \tag{11}$$

Following inequality (10), we further get

$$\begin{aligned}
r\delta \geq& \left(\frac{f_2}{f_0} - 1\right) \frac{m^2}{f_2\lambda + m^2} \frac{r(2-\delta)}{(m+1)^2} \\
\geq& \frac{f_2}{2f_0} \frac{m^2}{\lambda + \lambda} \frac{r}{4m^2} \\
\geq& \frac{r}{16 f_0 \lambda},
\end{aligned}$$

which is

$$\lambda \geq \frac{\delta^{-1}}{16 f_0} \tag{12}$$

For fixed ϵ,

$$f_0 \leq \frac{\epsilon}{U^2} + \frac{1}{TU^2} \leq \frac{\epsilon}{U^2} + \lambda^{-1},$$

and further

$$\begin{aligned} \lambda &\geq \frac{\delta^{-1}}{16(\frac{\epsilon}{U^2} + \lambda^{-1})} \\ \Rightarrow \lambda &\geq \frac{U^2}{\epsilon}\left(\frac{\delta^{-1}}{16} - 1\right) \geq \frac{U^2 \delta^{-1}}{32\epsilon} \end{aligned}$$

which shows that

$$T = \Theta(\lambda) = \Omega(U^2 \delta^{-1} \epsilon^{-1}).$$

When $\epsilon = \frac{1}{T}$,

$$f_0 \leq \frac{1}{U^2} \frac{\sum_{t=1}^{T} \frac{1}{t}}{T} \leq \frac{2\log(T)}{U^2 T}$$

and further

$$\begin{aligned} \lambda &\geq \frac{\delta^{-1}}{16 \frac{2\log(T)}{U^2 T}} \\ \Rightarrow \log(T) &\geq \frac{U^2 \sigma_e^2}{32 \sigma_w^2 \delta} \end{aligned}$$

which shows that

$$T = \exp(\Omega(U^2 \delta^{-1})).$$

□

2.4 Proof for Theorem 1.4 (ϵ-greedy explorationstructured exploration)

Theorem 1.4 *In the Climb Game $G_f(2, 0, U)$, given the quadratic joint Q function form $q(\mathbf{x}, \mathbf{y}; \mathbf{W}, \mathbf{b}, \mathbf{c}, d)$ and a Gaussian prior $p(\mathbf{W}) = \mathcal{N}(\mathbf{W}; 0, \sigma_w^2 I)$, under structured exploration $p_e^{(t)}(i, j) = U^{-1}[\mathbb{1}_{i=j}]$, $q_{\mathcal{J}^{(T)}}(\mathbf{W}, \mathbf{b}, \mathbf{c}, d)$ is equivalently optimal at step $T = O(1)$.*

Proof. It is easy to verify that $\mathbf{W} = \mathbf{c} = 0, \mathbf{b} = (1, 0, ..., 0)^\top, d = 0$ is the learned parameter that maximizes both the prior of $\mathbf{W}$ and the likelihood of prediction error at any step T. This parameter configuration directly gives the joint Q function that is equivalently optimal.

□