A. Appendix

A.1. Trajectory Distribution Induced by Logistic Stochastic Best Response Equilibrium

Let $\{\boldsymbol{\pi}_{-i}^t(\boldsymbol{a}_{-i}^t|s^t)\}_{t=1}^T$ denote other agents' marginal LSBRE policies, and $\{\hat{\pi}_i^t(a_i^t|\boldsymbol{a}_{-i}^t, s^t)\}_{t=1}^T$ denote agent i's conditional policy. With chain rule, the induced trajectory distribution is given by:

$$\hat{p}(\tau) = \left[\eta(s^1) \cdot \prod_{t=1}^T P(s^{t+1}|s^t, \boldsymbol{a}^t) \cdot \boldsymbol{\pi}_{-i}^t(\boldsymbol{a}_{-i}^t|s^t)\right] \cdot \prod_{t=1}^T \hat{\pi}_i^t(a_i^t|\boldsymbol{a}_{-i}^t, s^t) \tag{14}$$

Suppose the desired distribution is given by:

$$p(\tau) \propto \left[\eta(s^1) \cdot \prod_{t=1}^T P(s^{t+1}|s^t, \boldsymbol{a}^t) \cdot \boldsymbol{\pi}_{-i}^t(\boldsymbol{a}_{-i}^t|s^t)\right] \cdot \exp\left(\sum_{t=1}^T r_i(s^t, a_i^t, \boldsymbol{a}_{-i}^t)\right) \tag{15}$$

Now we will shown that the optimal solution to the following optimization problem correspond to the LSBRE conditional policies:

$$\min_{\hat{\pi}_i^{1:T}} D_{\mathrm{KL}}(\hat{p}(\tau)||p(\tau)) \tag{16}$$

The optimization problem in Equation (16) is equivalent to (the partition function of the desired distribution is a constant with respect to optimized policies):

$$\begin{aligned} \max_{\hat{\pi}_i^{1:T}} \; & \mathbb{E}_{\tau\sim\hat{p}(\tau)}\Bigg[\log\eta(s^1) + \sum_{t=1}^T(\log P(s^{t+1}|s^t, \boldsymbol{a}^t) + \log\boldsymbol{\pi}_{-i}^t(\boldsymbol{a}_{-i}^t|s^t) + r_i(s^t, \boldsymbol{a}^t)) - \\ & \qquad \log\eta(s^1) - \sum_{t=1}^T(\log P(s^{t+1}|s^t, \boldsymbol{a}^t) + \log\boldsymbol{\pi}_{-i}^t(\boldsymbol{a}_{-i}^t|s^t) + \log\hat{\pi}_i^t(a_i^t|\boldsymbol{a}_{-i}^t, s^t))\Bigg] \\ = \; & \mathbb{E}_{\tau\sim\hat{p}(\tau)}\left[\sum_{t=1}^T r_i(s^t, \boldsymbol{a}^t) - \log\hat{\pi}_i^t(a_i^t|\boldsymbol{a}_{-i}^t, s^t)\right] = \sum_{t=1}^T \mathbb{E}_{(s^t,\boldsymbol{a}^t)\sim\hat{p}(s^t,\boldsymbol{a}^t)}[r_i(s^t, \boldsymbol{a}^t) - \log\hat{\pi}_i^t(a_i^t|\boldsymbol{a}_{-i}^t, s^t)] \end{aligned} \tag{17}$$

To maximize this objective, we can use a dynamic programming procedure. Let us first consider the base case of optimizing $\hat{\pi}_i^T(a_i^T|\boldsymbol{a}_{-i}^T, s^T)$:

$$\begin{aligned} & \mathbb{E}_{(s^T,\boldsymbol{a}^T)\sim\hat{p}(s^T,\boldsymbol{a}^T)}[r_i(s^T, \boldsymbol{a}^T) - \log\hat{\pi}_i^T(a_i^T|\boldsymbol{a}_{-i}^T)] = \\ & \mathbb{E}_{s^T\sim\hat{p}(s^T),\boldsymbol{a}_{-i}^T\sim\boldsymbol{\pi}_{-i}^T(\cdot|s^T)}\left[-D_{\mathrm{KL}}\left(\hat{\pi}_i^T(a_i^T|\boldsymbol{a}_{-i}^T, s^T)||\frac{\exp(r_i(s^T, a_i^T, \boldsymbol{a}_{-i}^T))}{\exp(V_i(s^T, \boldsymbol{a}_{-i}^T))}\right) + V_i(s^T, \boldsymbol{a}_{-i}^T)\right] \end{aligned} \tag{18}$$

where $\exp(V_i(s^T, \boldsymbol{a}_{-i}^T))$ is the partition function and $V_i(s^T, \boldsymbol{a}_{-i}^T) = \log\sum_{a_i'}\exp(r_i(s^T, a_i', \boldsymbol{a}_{-i}^T))$. The optimal policy is given by:

$$\pi_i^T(a_i^T|\boldsymbol{a}_{-i}^T, s^T) = \exp(r_i(s^T, a_i^T, \boldsymbol{a}_{-i}^T) - V_i(s^T, \boldsymbol{a}_{-i}^T)) \tag{19}$$

With the optimal policy in Equation (19), Equation (18) is equivalent to (with the KL divergence being zero):

$$\mathbb{E}_{(s^T,\boldsymbol{a}^T)\sim\hat{p}(s^T,\boldsymbol{a}^T)}[r_i(s^T, \boldsymbol{a}^T) - \log\hat{\pi}_i^T(a_i^T|\boldsymbol{a}_{-i}^T)] = \mathbb{E}_{s^T\sim\hat{p}(s^T),\boldsymbol{a}_{-i}^T\sim\boldsymbol{\pi}_{-i}^T(\cdot|s^T)}[V_i(s^T, \boldsymbol{a}_{-i}^T)] \tag{20}$$

Then recursively, for a given time step t, $\hat{\pi}_i^t(a_i^t|\boldsymbol{a}_{-i}^t, s^t)$ must maximize:

$$\mathbb{E}_{(s^t,\boldsymbol{a}^t)\sim\hat{p}(s^t,\boldsymbol{a}^t)}\left[r_i(s^t, \boldsymbol{a}^t) - \log\hat{\pi}_i^t(a_i^t|\boldsymbol{a}_{-i}^t) + \mathbb{E}_{s^{t+1}\sim P(\cdot|s^t,\boldsymbol{a}^t),\boldsymbol{a}_{-i}^{t+1}\sim\boldsymbol{\pi}_{-i}^{t+1}(\cdot|s^{t+1})}[V_i^{\boldsymbol{\pi}^{t+2:T}}(s^{t+1}, \boldsymbol{a}_{-i}^{t+1})]\right] = \tag{21}$$

$$\mathbb{E}_{s^t\sim\hat{p}(s^t),\boldsymbol{a}_{-i}^t\sim\boldsymbol{\pi}_{-i}^t(\cdot|s^t)}\left[-D_{\mathrm{KL}}\left(\hat{\pi}_i^t(a_i^t|\boldsymbol{a}_{-i}^t, s^t)||\frac{\exp(Q_i^{\boldsymbol{\pi}^{t+1:T}}(s^t, a_i^t, \boldsymbol{a}_{-i}^t))}{\exp(V_i^{\boldsymbol{\pi}^{t+1:T}}(s^t, \boldsymbol{a}_{-i}^t))}\right) + V_i^{\boldsymbol{\pi}^{t+1:T}}(s^t, \boldsymbol{a}_{-i}^t)\right] \tag{22}$$

where we define:

$$Q_i^{\boldsymbol{\pi}^{t+1:T}}(s^t, \boldsymbol{a}^t) = r_i(s^t, \boldsymbol{a}^t) + \mathbb{E}_{s^{t+1}\sim p(\cdot|s^t,\boldsymbol{a}^t)}\left[\mathcal{H}(\pi_i^{t+1}(\cdot|s^{t+1})) + \mathbb{E}_{\boldsymbol{a}_{-i}^{t+1}\sim\boldsymbol{\pi}_{-i}^{t+1}(\cdot|s^{t+1})}[V_i(s^{t+1}, \boldsymbol{a}_{-i}^{t+1})]\right] \tag{23}$$

$$V_i^{\boldsymbol{\pi}^{t+1:T}}(s^t, \boldsymbol{a}_{-i}^t) = \log \sum_{a_i'} \exp(Q_i^{\boldsymbol{\pi}^{t+1:T}}(s^t, a_i', \boldsymbol{a}_{-i}^t)) \tag{24}$$

The optimal policy to Equation (22) is given by:

$$\pi_i^t(a_i^t|\boldsymbol{a}_{-i}^t, s^t) = \exp(Q_i^{\boldsymbol{\pi}^{t+1:T}}(s^t, \boldsymbol{a}^t) - V_i^{\boldsymbol{\pi}^{t+1:T}}(s^t, \boldsymbol{a}_{-i}^t)) \tag{25}$$

which is exactly the set of conditional distributions used to produce LSBRE (Definition 2).

A.2. Maximum Pseudolikelihood Estimation for LSBRE

Theorem 2 strictly follows the asymptotic consistency property of maximum pseudolikelihood estimation (Lehmann & Casella, 2006; Dawid & Musio, 2014). For simplicity, we will show the proof for normal form games and similar to Appendix A.1, the extension to Markov games can be proved by induction.

Consider a normal form game with N players and reward functions $\{r_i(\boldsymbol{a}; \omega_i)\}_{i=1}^N$. Suppose the expert demonstrations $\mathcal{D} = \{(a_1, \ldots, a_N)^m\}_{m=1}^M$ are generated by $\boldsymbol{\pi}(\boldsymbol{a}; \boldsymbol{\omega}^*)$, where $\boldsymbol{\omega}^*$ denotes the true value of the parameters. The pseudolikelihood objective we want to maximize is given by:

$$\begin{aligned}
\ell_{\mathrm{PL}}(\boldsymbol{\omega}) &= \frac{1}{M}\sum_{m=1}^M \sum_{i=1}^N \log \pi_i(a_i^m|\boldsymbol{a}_{-i}^m; \omega_i) = \frac{1}{M}\sum_{m=1}^M \sum_{i=1}^N \log \frac{\exp(r_i(a_i^m, \boldsymbol{a}_{-i}^m; \omega_i))}{\sum_{a_i'} \exp(r_i(a_i', \boldsymbol{a}_{-i}^m; \omega_i))} && (26)\\
&= \frac{1}{M}\sum_{m=1}^M \sum_{i=1}^N r_i(a_i^m, \boldsymbol{a}_{-i}^m; \omega_i) - \frac{1}{M}\sum_{m=1}^M \sum_{i=1}^N \log Z(\boldsymbol{a}_{-i}^m; \omega_i) && (27)\\
&= \sum_{i=1}^N \sum_{\boldsymbol{a}} p_{\mathcal{D}}(\boldsymbol{a}) r_i(a_i, \boldsymbol{a}_{-i}; \omega_i) - \sum_{i=1}^N \sum_{\boldsymbol{a}_{-i}} p_{\mathcal{D}}(\boldsymbol{a}_{-i}) \log Z(\boldsymbol{a}_{-i}; \omega_i) && (28)
\end{aligned}$$

where $p_{\mathcal{D}}$ is the empirical data distribution and $Z(\boldsymbol{a}_{-i}; \omega_i)$ is the partition function.

Take derivatives of $\ell_{\mathrm{PL}}(\boldsymbol{\omega})$:

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\omega}} \ell_{\mathrm{PL}}(\boldsymbol{\omega}) &= \sum_{i=1}^N \sum_{\boldsymbol{a}} p_{\mathcal{D}}(\boldsymbol{a}) \frac{\partial}{\partial \boldsymbol{\omega}} r_i(a_i, \boldsymbol{a}_{-i}; \omega_i) - \sum_{i=1}^N \sum_{\boldsymbol{a}_{-i}} p_{\mathcal{D}}(\boldsymbol{a}_{-i}) \frac{1}{Z(\boldsymbol{a}_{-i}; \omega_i)} \frac{\partial}{\partial \boldsymbol{\omega}} Z(\boldsymbol{a}_{-i}; \omega_i) && (29)\\
&= \sum_{i=1}^N \sum_{\boldsymbol{a}} p_{\mathcal{D}}(\boldsymbol{a}) \frac{\partial}{\partial \boldsymbol{\omega}} r_i(a_i, \boldsymbol{a}_{-i}; \omega_i) - \sum_{i=1}^N \sum_{\boldsymbol{a}_{-i}} p_{\mathcal{D}}(\boldsymbol{a}_{-i}) \sum_{a_i} \frac{\exp(r_i(a_i, \boldsymbol{a}_{-i}; \omega_i))}{Z(\boldsymbol{a}_{-i}; \omega_i)} \frac{\partial}{\partial \boldsymbol{\omega}} r_i(a_i, \boldsymbol{a}_{-i}; \omega_i) && (30)\\
&= \sum_{i=1}^N \sum_{\boldsymbol{a}} p_{\mathcal{D}}(\boldsymbol{a}) \frac{\partial}{\partial \boldsymbol{\omega}} r_i(a_i, \boldsymbol{a}_{-i}; \omega_i) - \sum_{i=1}^N \sum_{\boldsymbol{a}_{-i}} p_{\mathcal{D}}(\boldsymbol{a}_{-i}) \sum_{a_i} \pi_i(a_i|\boldsymbol{a}_{-i}; \omega_i) \frac{\partial}{\partial \boldsymbol{\omega}} r_i(a_i, \boldsymbol{a}_{-i}; \omega_i) && (31)
\end{aligned}$$

When the sample size $m \to \infty$, Equation (31) is equivalent to:

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\omega}} \ell_{\mathrm{PL}}(\boldsymbol{\omega}) &= \sum_{i=1}^N \sum_{\boldsymbol{a}} p(\boldsymbol{a}; \boldsymbol{\omega}^*) \frac{\partial}{\partial \boldsymbol{\omega}} r_i(a_i, \boldsymbol{a}_{-i}; \omega_i) - \sum_{i=1}^N \sum_{\boldsymbol{a}_{-i}} p(\boldsymbol{a}_{-i}; \boldsymbol{\omega}^*) \sum_{a_i} \pi_i(a_i|\boldsymbol{a}_{-i}; \omega_i) \frac{\partial}{\partial \boldsymbol{\omega}} r_i(a_i, \boldsymbol{a}_{-i}; \omega_i) && (32)\\
&= \sum_{i=1}^N \sum_{\boldsymbol{a}_{-i}} p(\boldsymbol{a}_{-i}; \boldsymbol{\omega}^*) \sum_{a_i} (p(a_i|\boldsymbol{a}_{-i}; \boldsymbol{\omega}^*) - \pi_i(a_i|\boldsymbol{a}_{-i}; \omega_i)) \frac{\partial}{\partial \boldsymbol{\omega}} r_i(a_i, \boldsymbol{a}_{-i}; \omega_i) && (33)
\end{aligned}$$

When $\boldsymbol{\omega} = \boldsymbol{\omega}^*$, the gradients in Equation (33) will be zero.