Acknowledgments

This research was supported by Toyota Research Institute, NSF (#1651565, #1522054, #1733686), ONR (N00014-19-1-2145), AFOSR (FA9550-19-1-0024), Amazon AWS, and Qualcomm.

References

Abbeel, P. and Ng, A. Y. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the twenty-first international conference on Machine learning*, pp. 1, 2004.

Amodei, D. and Clark, J. Faulty reward functions in the wild, 2016.

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., and Mané, D. Concrete problems in AI safety. *arXiv preprint arXiv:1606. 06565*, 2016.

Arnold, B. C. and Press, S. J. Compatible conditional distributions. *Journal of the American Statistical Association*, 84(405):152–156, 1989.

Aumann, R. J. Subjectivity and correlation in randomized strategies. *Journal of mathematical Economics*, 1(1): 67–96, 1974.

Aumann, R. J. Correlated equilibrium as an expression of bayesian rationality. *Econometrica: Journal of the Econometric Society*, pp. 1–18, 1987.

Barrett, S., Rosenfeld, A., Kraus, S., and Stone, P. Making friends on the fly: Cooperating with new teammates. *Artificial Intelligence*, 242:132–171, 2017.

Besag, J. Statistical analysis of non-lattice data. *The statistician*, pp. 179–195, 1975.

Bogert, K. and Doshi, P. Multi-robot inverse reinforcement learning under occlusion with interactions. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, pp. 173–180, 2014.

Chen, S.-H. and Ip, E. H. Behaviour of the gibbs sampler when conditional distributions are potentially incompatible. *Journal of statistical computation and simulation*, 85(16):3266–3275, 2015.

Chen, S.-H., Ip, E. H., and Wang, Y. J. Gibbs ensembles for nearly compatible and incompatible conditional models. *Computational statistics & data analysis*, 55(4):1760–1769, 2011.

Dawid, A. P. and Musio, M. Theory and applications of proper scoring rules. *Metron*, 72(2):169–183, 2014.

Devlin, S. and Kudenko, D. Theoretical considerations of potential-based reward shaping for multi-agent systems. In *The 10th International Conference on Autonomous Agents and Multiagent Systems - Volume 1*, AAMAS '11, pp. 225–232, Richland, SC, 2011. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 9780982657157, 9780982657157.

Finn, C., Christiano, P., Abbeel, P., and Levine, S. A connection between generative adversarial networks, inverse reinforcement learning, and energy-based models. *arXiv preprint arXiv:1611.03852*, 2016a.

Finn, C., Levine, S., and Abbeel, P. Guided cost learning: Deep inverse optimal control via policy optimization. In *International Conference on Machine Learning*, pp. 49–58, June 2016b.

Fu, J., Luo, K., and Levine, S. Learning robust rewards with adversarial inverse reinforcement learning. *arXiv preprint arXiv:1710.11248*, 2017.

Gandhi, A. The stochastic response dynamic: A new approach to learning and computing equilibrium in continuous games. *Technical Report*, 2012.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *Advances in neural information processing systems*, pp. 2672–2680, 2014.

Gordon, G. J., Greenwald, A., and Marks, C. No-regret learning in convex games. In *Proceedings of the 25th international conference on Machine learning*, pp. 360–367. ACM, 2008.

Gu, S., Holly, E., Lillicrap, T., and Levine, S. Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates. In *Robotics and Automation (ICRA), 2017 IEEE International Conference on*, pp. 3389–3396. IEEE, 2017.

Hadfield-Menell, D., Milli, S., Abbeel, P., Russell, S. J., and Dragan, A. Inverse reward design. In *Advances in Neural Information Processing Systems*, pp. 6765–6774, 2017.

Hart, S. and Mas-Colell, A. A simple adaptive procedure leading to correlated equilibrium. *Econometrica*, 68(5): 1127–1150, 2000.

Hastings, W. K. Monte carlo sampling methods using markov chains and their applications. 1970.

Heckerman, D., Chickering, D. M., Meek, C., Rounthwaite, R., and Kadie, C. Dependency networks for inference, collaborative filtering, and data visualization. *Journal of Machine Learning Research*, 1(Oct):49–75, 2000.

Ho, J. and Ermon, S. Generative adversarial imitation learning. In *Advances in Neural Information Processing Systems*, pp. 4565–4573, 2016.

Ho, J., Gupta, J., and Ermon, S. Model-free imitation learning with policy optimization. In *International Conference on Machine Learning*, pp. 2760–2769, 2016.

Hu, J., Wellman, M. P., and Others. Multiagent reinforcement learning: theoretical framework and an algorithm. In *ICML*, volume 98, pp. 242–250, 1998.

Kalakrishnan, M., Pastor, P., Righetti, L., and Schaal, S. Learning objective functions for manipulation. In *Robotics and Automation (ICRA), 2013 IEEE International Conference on*, pp. 1331–1336, 2013.

Le, H. M., Yue, Y., and Carr, P. Coordinated Multi-Agent imitation learning. *arXiv preprint arXiv:1703.03121*, March 2017.

Lehmann, E. L. and Casella, G. *Theory of point estimation*. Springer Science & Business Media, 2006.

Leibo, J. Z., Zambaldi, V., Lanctot, M., Marecki, J., and Graepel, T. Multi-agent reinforcement learning in sequential social dilemmas. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*, pp. 464–473, 2017.

Levine, S. Reinforcement learning and control as probabilistic inference: Tutorial and review. *arXiv preprint arXiv:1805. 00909*, 2018.

Levine, S., Finn, C., Darrell, T., and Abbeel, P. End-to-end training of deep visuomotor policies. *The Journal of Machine Learning Research*, 17(1):1334–1373, 2016.

Li, Y., Song, J., and Ermon, S. InfoGAIL: Interpretable imitation learning from visual demonstrations. *arXiv preprint arXiv:1703. 08840*, 2017.

Lin, X., Beling, P. A., and Cogill, R. Multi-agent inverse reinforcement learning for zero-sum games. *arXiv preprint arXiv:1403. 6508*, 2014.

Lin, X., Adams, S. C., and Beling, P. A. Multi-agent inverse reinforcement learning for general-sum stochastic games. *arXiv preprint arXiv:1806.09795*, 2018.

Littman, M. L. Markov games as a framework for multi-agent reinforcement learning. In *Proceedings of the eleventh international conference on machine learning*, volume 157, pp. 157–163, 1994.

Lowe, R., Wu, Y., Tamar, A., Harb, J., Abbeel, P., and Mordatch, I. Multi-Agent Actor-Critic for mixed Cooperative-Competitive environments. *arXiv preprint arXiv:1706.02275*, June 2017.

Matignon, L., Jeanpierre, L., Mouaddib, A.-I., and Others. Coordinated Multi-Robot exploration under communication constraints using decentralized markov decision processes. In *AAAI*, 2012.

McKelvey, R. D. and Palfrey, T. R. Quantal response equilibria for normal form games. *Games and economic behavior*, 10(1):6–38, 1995.

McKelvey, R. D. and Palfrey, T. R. Quantal response equilibria for extensive form games. *Experimental economics*, 1(1):9–41, 1998.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540): 529, 2015.

Natarajan, S., Kunapuli, G., Judah, K., Tadepalli, P., Kersting, K., and Shavlik, J. Multi-agent inverse reinforcement learning. In *Machine Learning and Applications (ICMLA), 2010 Ninth International Conference on*, pp. 395–400, 2010.

Ng, A. Y., Harada, D., and Russell, S. Policy invariance under reward transformations: Theory and application to reward shaping. In *ICML*, volume 99, pp. 278–287, 1999.

Ng, A. Y., Russell, S. J., et al. Algorithms for inverse reinforcement learning. In *Icml*, pp. 663–670, 2000.

Nisan, N., Schapira, M., Valiant, G., and Zohar, A. Best-response mechanisms. In *ICS*, pp. 155–165, 2011.

Peng, P., Yuan, Q., Wen, Y., Yang, Y., Tang, Z., Long, H., and Wang, J. Multiagent Bidirectionally-Coordinated nets for learning to play StarCraft combat games. *arXiv preprint arXiv:1703. 10069*, 2017.

Pomerleau, D. A. Efficient training of artificial neural networks for autonomous navigation. *Neural computation*, 3(1):88–97, 1991. ISSN 0899-7667.

Reddy, T. S., Gopikrishna, V., Zaruba, G., and Huber, M. Inverse reinforcement learning for decentralized non-cooperative multiagent systems. In *Systems, Man, and Cybernetics (SMC), 2012 IEEE International Conference on*, pp. 1930–1935, 2012.

Russell, S. Learning agents for uncertain environments. In *Proceedings of the eleventh annual conference on Computational learning theory*, pp. 101–103. ACM, 1998.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484, 2016.

Song, J., Ren, H., Sadigh, D., and Ermon, S. Multi-agent generative adversarial imitation learning. 2018.

Šošić, A., KhudaBukhsh, W. R., Zoubir, A. M., and Koeppl, H. Inverse reinforcement learning in swarm systems. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*, pp. 1413–1421. International Foundation for Autonomous Agents and Multiagent Systems, 2017.

Waugh, K., Ziebart, B. D., and Andrew Bagnell, J. Computational rationalization: The inverse equilibrium problem. *arXiv preprint arXiv:1308.3506*, August 2013.

Wu, Y., Mansimov, E., Liao, S., Grosse, R., and Ba, J. Scalable trust-region method for deep reinforcement learning using kronecker-factored approximation. *arXiv preprint arXiv:1708.05144*, August 2017.

Yu, L., Zhang, W., Wang, J., and Yu, Y. Seqgan: Sequence generative adversarial nets with policy gradient. In *AAAI*, pp. 2852–2858, 2017.

Ziebart, B. D. Modeling purposeful adaptive behavior with the principle of maximum causal entropy. 2010.

Ziebart, B. D., Maas, A. L., Bagnell, J. A., and Dey, A. K. Maximum entropy inverse reinforcement learning. In *AAAI*, volume 8, pp. 1433–1438, 2008.

Ziebart, B. D., Bagnell, J. A., and Dey, A. K. Maximum causal entropy correlated equilibria for markov games. In *The 10th International Conference on Autonomous Agents and Multiagent Systems - Volume 1*, AAMAS '11, pp. 207–214, Richland, SC, 2011. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 9780982657157, 9780982657157.

Zoph, B. and Le, Q. V. Neural architecture search with reinforcement learning. *arXiv preprint arXiv:1611.01578*, 2016.

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:

$$\ell_{\mathrm{PL}}(\boldsymbol{\omega}) = \frac{1}{M}\sum_{m=1}^M\sum_{i=1}^N\log\pi_i(a_i^m|\boldsymbol{a}_{-i}^m;\omega_i) = \frac{1}{M}\sum_{m=1}^M\sum_{i=1}^N\log\frac{\exp(r_i(a_i^m, \boldsymbol{a}_{-i}^m;\omega_i))}{\sum_{a_i'}\exp(r_i(a_i', \boldsymbol{a}_{-i}^m;\omega_i))} \tag{26}$$

$$= \frac{1}{M}\sum_{m=1}^M\sum_{i=1}^N r_i(a_i^m, \boldsymbol{a}_{-i}^m;\omega_i) - \frac{1}{M}\sum_{m=1}^M\sum_{i=1}^N\log Z(\boldsymbol{a}_{-i}^m;\omega_i) \tag{27}$$

$$= \sum_{i=1}^N\sum_{\boldsymbol{a}} p_{\mathcal{D}}(\boldsymbol{a})r_i(a_i, \boldsymbol{a}_{-i};\omega_i) - \sum_{i=1}^N\sum_{\boldsymbol{a}_{-i}} p_{\mathcal{D}}(\boldsymbol{a}_{-i})\log Z(\boldsymbol{a}_{-i};\omega_i) \tag{28}$$

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

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

$$\frac{\partial}{\partial\boldsymbol{\omega}}\ell_{\mathrm{PL}}(\boldsymbol{\omega}) = \sum_{i=1}^N\sum_{\boldsymbol{a}} p_{\mathcal{D}}(\boldsymbol{a})\frac{\partial}{\partial\boldsymbol{\omega}}r_i(a_i, \boldsymbol{a}_{-i};\omega_i) - \sum_{i=1}^N\sum_{\boldsymbol{a}_{-i}} p_{\mathcal{D}}(\boldsymbol{a}_{-i})\frac{1}{Z(\boldsymbol{a}_{-i};\omega_i)}\frac{\partial}{\partial\boldsymbol{\omega}}Z(\boldsymbol{a}_{-i};\omega_i) \tag{29}$$

$$= \sum_{i=1}^N\sum_{\boldsymbol{a}} p_{\mathcal{D}}(\boldsymbol{a})\frac{\partial}{\partial\boldsymbol{\omega}}r_i(a_i, \boldsymbol{a}_{-i};\omega_i) - \sum_{i=1}^N\sum_{\boldsymbol{a}_{-i}} p_{\mathcal{D}}(\boldsymbol{a}_{-i})\sum_{a_i}\frac{\exp(r_i(a_i, \boldsymbol{a}_{-i};\omega_i))}{Z(\boldsymbol{a}_{-i};\omega_i)}\frac{\partial}{\partial\boldsymbol{\omega}}r_i(a_i, \boldsymbol{a}_{-i};\omega_i) \tag{30}$$

$$= \sum_{i=1}^N\sum_{\boldsymbol{a}} p_{\mathcal{D}}(\boldsymbol{a})\frac{\partial}{\partial\boldsymbol{\omega}}r_i(a_i, \boldsymbol{a}_{-i};\omega_i) - \sum_{i=1}^N\sum_{\boldsymbol{a}_{-i}} p_{\mathcal{D}}(\boldsymbol{a}_{-i})\sum_{a_i}\pi_i(a_i|\boldsymbol{a}_{-i};\omega_i)\frac{\partial}{\partial\boldsymbol{\omega}}r_i(a_i, \boldsymbol{a}_{-i};\omega_i) \tag{31}$$

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

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

$$= \sum_{i=1}^N\sum_{\boldsymbol{a}_{-i}} p(\boldsymbol{a}_{-i};\boldsymbol{\omega}^*)\sum_{a_i}(p(a_i|\boldsymbol{a}_{-i};\boldsymbol{\omega}^*) - \pi_i(a_i|\boldsymbol{a}_{-i};\omega_i))\frac{\partial}{\partial\boldsymbol{\omega}}r_i(a_i, \boldsymbol{a}_{-i};\omega_i) \tag{33}$$

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