Global Convergence of Multi-Agent Policy Gradient in Markov Potential Games

Stefanos Leonardos
SUTD

Will Overman
UC Irvine

Ioannis Panageas
UC Irvine

Georgios Piliouras
SUTD

Abstract

Potential games are arguably one of the most important and widely studied classes of normal form games. They define the archetypal setting of multi-agent coordination as all agent utilities are perfectly aligned with each other via a common potential function. Can this intuitive framework be transplanted in the setting of Markov Games? What are the similarities and differences between multi-agent coordination with and without state dependence? We present a novel definition of Markov Potential Games (MPG) that generalizes prior attempts at capturing complex stateful multi-agent coordination. Counter-intuitively, insights from normal-form potential games do not carry over as MPGs can consist of settings where state-games can be zero-sum games. In the opposite direction, Markov games where every state-game is a potential game are not necessarily MPGs. Nevertheless, MPGs showcase standard desirable properties such as the existence of deterministic Nash policies. In our main technical result, we prove (polynomially fast in the approximation error) convergence of independent policy gradient to Nash policies by adapting recent gradient dominance property arguments developed for single agent MDPs to multi-agent learning settings.

1. Extended Abstract

Reinforcement learning (RL) has been a fundamental driver of numerous recent advances in Artificial Intelligence (AI) applications that range from super-human performance in competitive game-playing (Silver et al., 2016; 2018; Brown and Sandholm, 2019) and strategic decision-making in multiple tasks (Mnih et al., 2015; OpenAI, 2018; Vinyals et al., 2019) to robotics, autonomous-driving and cyber-physical systems (Busoniu et al., 2008; Zhang et al., 2019). A core ingredient for the success of single-agent RL systems, which are typically modelled as Markov Decision Processes (MDPs), is the guarantee of existence of stationary deterministic optimal policies (Bertsekas, 2000; Sutton and Barto,

2018). This allows for the design of efficient algorithms that provably converge towards the optimal policy (Agarwal et al., 2020). However, a majority of the above systems involve multi-agent interactions and despite the notable empirical advancements, there is a lack of understanding about the theoretical convergence guarantees of the existing multi-agent reinforcement learning (MARL) algorithms.

The main challenge when transitioning from single to multi-agent RL settings is the computation of *Nash policies*. A Nash policy for $n > 1$ agents is defined to be a profile of policies $(\pi_1^*, ..., \pi_n^*)$ so that by fixing the stationary policies of all agents but i, π_i^* is an optimal policy for the resulting single-agent MDP and this is true for all $1 \leq i \leq n$ [1] (see Definition 1). Note that in multi-agent settings, Nash policies *may not be unique* in principle.

A common approach for computing Nash policies in MDPs is the use of *policy gradient* methods. There has been significant progress in the analysis of policy gradient methods during the last couple of years, notably including the works of (Agarwal et al., 2020) (and references therein), but it has mainly concerned the single-agent case: the convergence properties of policy gradient in MARL remain poorly understood. Existing steps towards a theory for multi-agent settings involve the papers of (Daskalakis et al., 2020) who show convergence of *independent policy gradient* to the optimal policy, for two-agent zero-sum stochastic games, of (Wei et al., 2021) who improve the result of (Daskalakis et al., 2020) using optimistic policy gradient and of (Zhao et al., 2021) who study extensions of Natural Policy Gradient using function approximation. It is worth noting that the positive results of (Daskalakis et al., 2020; Wei et al., 2021) and (Zhao et al., 2021) depend on the fact that two-agent stochastic zero-sum games satisfy the "min-max equals max-min" property (Shapley, 1953) (even though the value-function landscape may not be convex-concave, which implies that Von Neumann's celebrated minimax theorem may not be applicable).

Model and Informal Statement of Results. While the previous works enhance our understanding in *competitive*

---

[1] Analogue of Nash equilibrium notion.

interactions, i.e., interactions in which gains can only come at the expense of others, MARL in *cooperative* settings remains largely under-explored and constitutes one of the current frontiers in AI research (Dafoe et al., 2020; Dafoe et al., 2021). Based on the above, our work is motivated by the following natural question:

Can we get (provably) fast convergence guarantees for multi-agent RL settings in which cooperation is desirable?

To address this question, we define and study a class of n-agent MDPs that naturally generalize normal form potential games (Monderer and Shapley, 1996), called *Markov Potential Games (MPGs)*. In words, a multi-agent MDP is a MPG as long as there exists a (state-dependent) real-valued potential function Φ so that if an agent i changes their policy (and the rest of the agents keep their policy unchanged), the difference in agent i's value/utility, V^i, is captured by the difference in the value of Φ (see Definition 2). Weighted and ordinal MPGs are defined similar to the normal form counterparts (see Remark 1).

Under our definition, we answer the above motivating question in the affirmative. In particular, we show that if every agent i independently runs (with simultaneous updates) policy gradient on his utility/value V^i, after $O(1/\epsilon^2)$ iterations, the system will reach an ϵ-approximate Nash policy (see informal Theorem 1.1 and formal Theorem 4.5). Moreover, we show the finite sample analogue, that is if every agent i independently runs (with simultaneous updates) stochastic policy gradient, then with high probability, the system will reach an ϵ-approximate Nash policy after $O(1/\epsilon^6)$ iterations.

Along the way, we prove several properties about the structure of MPGs and their Nash policies (see Theorem 1.2 and Section 3). Our results can be summarized in the following two Theorems.

Theorem 1.1 (Convergence of Policy Gradient (Informal)). *Consider a MPG with n agents and let $\epsilon > 0$. Suppose that each agent i runs independent policy gradient using direct parameterization on his policy and that the updates are simultaneous. Then, the learning dynamics reach an ϵ-Nash policy after $\mathcal{O}(1/\epsilon^2)$ iterations. Moreover, suppose that each agent i runs stochastic policy gradient using greedy parameterization (see (4)) on his policy and that the updates are simultaneous. Then the learning dynamics reach an ϵ-Nash policy after $\mathcal{O}(1/\epsilon^6)$ iterations.*

This result holds trivially for weighted MPGs and asymptotically also for ordinal MPGs, see Remark 4.

Theorem 1.2 (Structural Properties of MPGs). *The following facts are true for MPGs with n-agents:*

 a. *There always exists a Nash policy profile $(\pi_1^*, \ldots, \pi_n^*)$*
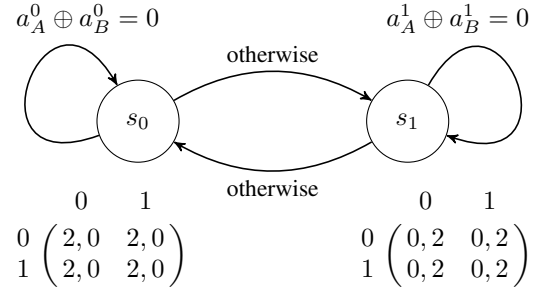

Figure 1. A MDP which is potential at every state but which not a MPG due to conflicting preferences over states. The agents' instantaneous rewards, $(R_A(s, \mathbf{a}), R_B(s, \mathbf{a}))$, are shown in matrix form below each state $s = 0, 1$.
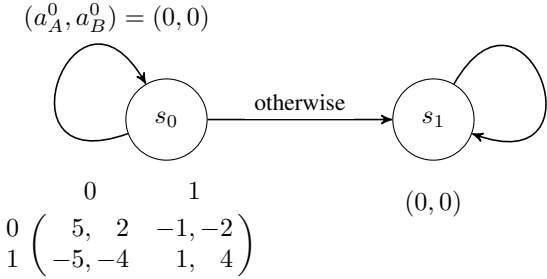

Figure 2. A MDP which is potential at every state and cooperative in nature but which is not a MPG. The action-dependent transitions do not allow the derivation of an exact potential function.

so that π_i^* is deterministic for each agent i (see Theorem 3.1).

 b. *We can construct MDPs for which each state is an underlying potential game but the MDPs are not MPGs. This can be true regardless of whether the whole MDP is competitive or cooperative in nature (see Figures 1 and 2, respectively). On the opposite side, we can construct MDPs that are MPGs but which include states that are purely competitive (i.e., zero-sum games), see Example 3.*

 c. *We provide sufficient conditions so that a MDP is a MPG. These include cases where each state is an underlying potential game and the transition probabilities are not affected by agents actions or the reward functions satisfy certain regularity conditions between different states (see conditions C1 and C2 in Proposition 3.2).*

Technical Overview. The first challenge in the proof of Theorem 1.1 is that multi-agent settings (MPGs) do not satisfy the gradient dominance property, which is an important part in the proof of convergence of policy gradient in single-
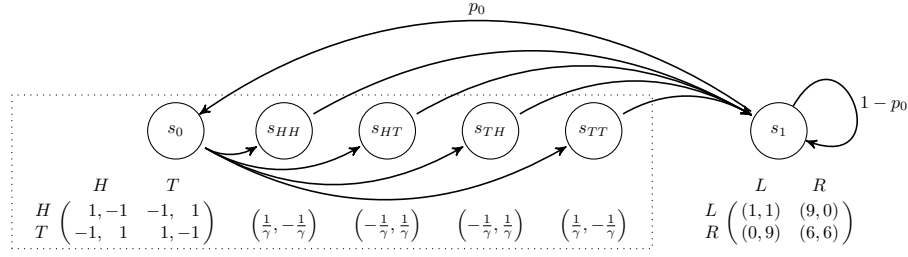
Figure 3. A 2-player MPG which is not potential at every state. The rewards in state s_1 form a potential game, whereas the rewards in s_0 do not. However, the states inside the dotted rectangle do form a potential game and this can be leveraged to show that the whole MPG is a potential game whenever p_0 does not depend on agents' actions.

agent settings (Agarwal et al., 2020). In particular, there is no uniqueness of optimal policies and as a result, there is not a properly defined notion of value in MPGs (in contrast to zero-sum stochastic games (Daskalakis et al., 2020)). On the positive side, we show that agent-wise (i.e., after fixing the policy of all agents but i), the value function, V^i, satisfies the gradient dominance property along the direction of π_i (policy of agent i). This can be leveraged to show that every *(approximate) stationary point* (Definition 4) of the potential function Φ is an *(approximate) Nash policy* (Lemma 4.2). As a result, convergence to an approximate Nash policy is established by showing that Φ is smooth and then applying *Projected Gradient Ascent* (PGA) on Φ. This step uses the rather well-known fact that (PGA) converges to ϵ-stationary points in $O(1/\epsilon^2)$ iterations for smooth functions. As a result, by applying PGA on the potential Φ, one gets an approximate Nash policy. Our convergence result then follows by showing that PGA on the potential function, Φ, generates the same dynamics as if each agent i runs independent PGA on their value function, V^i.

In the case that agents do not have access to exact gradients, we derive a similar result for finite samples. In this case, we apply *Projected Stochastic Gradient Ascent (PSGA)* on Φ which (as was the case for PGA) can be shown to be the same as when agents apply PSGA independently on their individual value functions. The key is to get an *unbiased sample* for the gradient of the value functions and prove that it has bounded variance (in terms of the parameters of the MPG). This comes from the discount factor, γ; in this case, $1 - \gamma$ can be interpreted as the probability to terminate the MDP at a particular state (and γ to continue). This can be used to show that a trajectory of the MDP is an unbiased sample for the gradient of the value functions. To guarantee that the estimate has bounded variance, we apply the approach of (Daskalakis et al., 2020) which requires that agents perform PSGA with α-greedy exploration (see (4)). The main idea is that this parameterization stays away from the boundary of the simplex throughout its trajectory.

Concerning our structural results, the main technical challenge is the dependence of state-transitions (in addition to

agents' rewards) on agents' actions. Our work in this part is mainly concerned with showing that the class of MPGs can be significantly larger than state based potential games but also that even simple coordination games may fail to satisfy the (exact) MPG property. Finally, concerning the existence of a deterministic Nash policies, the main challenge is (as in Theorem 1.1) the lack of a value in general multi-agent settings. As we show in the proof of Theorem 3.1, this issue can be still handled within the class of MPGs by constructing single-agent deviations (to deterministic optimal policies) which keep the value of the potential constant (at its global maximum). This process (which leads to a deterministic Nash policy profile) depends on the MPG property and does not generalize to arbitrary MARL settings.

Other works on MPGs. There are only a few papers in the recent literature that define and analyze MARL settings under the term *Markov Potential Games* using slight different definitions (see (Marden, 2012; Valcarcel Macua et al., 2018)).[2] These papers mainly focus on state-based potential MDPs (i.e., MPDs in which every state is a potential game) and require rather restrictive additional conditions, such as equality or monotonicity of the state-based potential functions, to address the computational challenge of finding Nash policies.[3] Our current results demonstrate the efficiency of simultaneous policy gradient as a to powerful method to find Nash policies even without additional restrictive assumptions on the state-based potential functions. Moreover, as mentioned in Theorem 1.2, the current definition also encompasses MDPs that are not necessarily potential at each state. To the best of our knowledge, the only (cooperative) MPGs that have been successfully addressed are the ones in which all agents receive the same value/utility (Wang and Sandholm, 2002).

Experimental Results We empirically studied the performance of policy gradient in a general class of MPGs that are congestion games at every state (cf. (Bistritz and Bambos,

---

[2] Ongoing works include (Mguni, 2020; Mguni et al., 2021).

[3] The relation of these conditions to the current work is discussed in more detail in Proposition 3.2 and Remark 2.
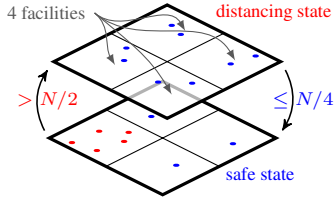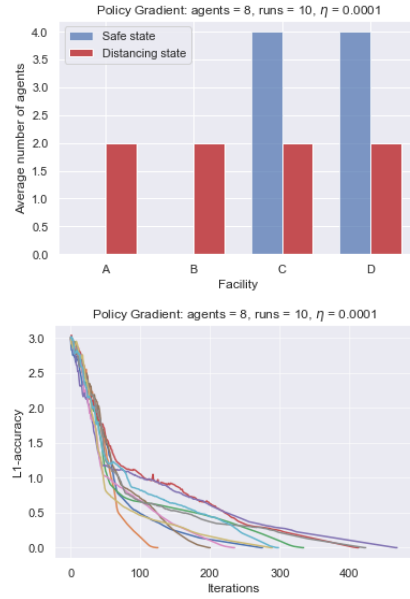
Figure 4. The 2-state MPG.

Figure 5. Policy gradient in the 2-state MPG with 8 agents of Figure 4. In all runs, the 8 agents learn one of the deterministic Nash policies that leads to the optimal distribution among states (left). Individual trajectories of the L1-accuracy and averages show fast convergence in all cases.

2020)). The setting of the current experiment is illustrated in Figure 4.

Experimental setup. There are 8 agents, 4 facilities and 2 states: a *safe* state and a *distancing* state. In both states, all agents prefer to be in the same facility with as many other agents as possible *(follow the crowd)* (Hassin and Haviv, 2003). In particular, the reward of each agent for being at facility $k = A, B, C, D$ is equal to a predefined positive weight w_k^{safe} times the number of agents at that facility. The weights satisfy $w_A^{\text{safe}} < w_B^{\text{safe}} < w_C^{\text{safe}} < w_D^{\text{safe}}$, i.e., facility D is the most preferable by all agents. If more than $4 = N/2$ agents find themselves in the same facility, then the game transitions to the distancing state. At that state, the reward structure remains the same, but the weights are reduced by a constant factor, i.e., $w_k^{\text{dist}} = w_k^{\text{safe}} - c$, where $c > 0$ is a (considerably large) constant. To return to the safe state, the agents need to achieve maximum distribution over the facilities, i.e., no more than $2 = N/4$ agents may be in the same facility.

To see that this MDP is a MPG, it suffices to check that every state is a potential game and that condition C2 (i.e., equality of individual dummy terms) of Proposition 3.2 is satisfied. The first claim is straightforward since at each state, the agents play a congestion game (Monderer and Shapley, 1996; Roughgarden, 2015). The second claim follows from the fact that the rewards of all agents in all facilities at the distancing state are shifted by the same constant amount, c.

Results. The top panel of Figure 5 shows that the agents learn the expected Nash profile in both states in all runs. The trajectories in the bottom panel depict the L1-accuracy in the policy space at each iteration which is defined as the average distance between the current policy and the final policy of all 8 agents, i.e., L1-accuracy $= \frac{1}{N} \sum_{i \in \mathcal{N}} |\pi_i - \pi_i^{\text{final}}| = \frac{1}{N} \sum_{i \in \mathcal{N}} \sum_s \sum_a |\pi_i(a \mid s) - \pi_i^{\text{final}}(a \mid s)|$.

Open questions. When it comes to online learning in normal form potential games, it is possible to prove that many naturally motivated dynamics converge to deterministic Nash equilibria with certain desirable stability properties for most initial conditions (Kleinberg et al., 2009; Panageas et al., 2018; Cohen et al., 2017; Mehta et al., 2015). To produce such equilibrium selection results, standard Lyapunov arguments do not suffice and one needs to apply more advanced techniques such as the Center-Stable-Manifold theorem (Lee et al., 2019). Studying such techniques in the context of MPGs is a fascinating direction for future work.

On the other hand, given the complexities of multi-agent, state-based environments, it is highly unlikely to expect that practical algorithms can always guarantee convergence to equilibrium. This is already the case even for the more restricted settings of normal-form games (Vlatakis-Gkaragkounis et al., 2020; Andrade et al., 2021). Nevertheless, deriving strong theoretical guarantees in the sense of cyclic/recurrent orbits, invariant functions (Mertikopoulos et al., 2018) or social welfare (Syrgkanis et al., 2015) in the context of exact, weighted or ordinal MPGs is another stimulating direction for future work. As a measurement of the inefficiency due to lack of coordination between agents, it would also be interesting to perform a Price of Anarchy type of analysis (Koutsoupias and Papadimitriou, 1999) as has been excessively done in the context of potential (congestion) games (e.g., (Roughgarden and Tardos, 2002)).

Finally, other natural directions for future work involve the study of policy gradient or variations thereof (such as Natural Policy Gradient) in MPGs under different policy parametrizations, cf. (Agarwal et al., 2020), or the study of settings that fruitfully combine tools from both cooperative and competitive settings (as in (Daskalakis et al., 2020; Wei et al., 2021; Zhao et al., 2021)) that have (up to now) produced results in orthogonal directions.

References

A. Agarwal, S. M. Kakade, J. D. Lee, and G. Mahajan. Optimality and Approximation with Policy Gradient Methods in Markov Decision Processes. In J. Abernethy and S. Agarwal, editors, *Proceedings of 33rd Conference on Learning Theory*, volume 125 of *PMLR*, pages 64–66, 2020. URL http://proceedings.mlr.press/v125/agarwal20a.html.

Gabriel P Andrade, Rafael Frongillo, and Georgios Piliouras. Learning in matrix games can be arbitrarily complex. *arXiv preprint arXiv:2103.03405*, 2021.

Dimitri P. Bertsekas. *Dynamic Programming and Optimal Control*. Athena Scientific, 2nd edition, 2000. ISBN 1886529094.

I. Bistritz and N. Bambos. Cooperative multi-player bandit optimization. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 2016–2027. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/15ae3b9d6286f1b2a489ea4f3f4abaed-Paper.pdf.

Noam Brown and Tuomas Sandholm. Superhuman ai for multiplayer poker. *Science*, 365(6456):885–890, 2019. ISSN 0036-8075. doi: 10.1126/science.aay2400.

Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Found. Trends Mach. Learn.*, 8(3-4):231–357, 2015.

Lucian Busoniu, Robert Babuska, and Bart De Schutter. A comprehensive survey of multiagent reinforcement learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 38(2):156–172, 2008. doi: 10.1109/TSMCC.2007.913919.

Johanne Cohen, Amélie Héliou, and Panayotis Mertikopoulos. Learning with bandit feedback in potential games. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6372–6381, 2017.

A. Dafoe, E. Hughes, Y. Bachrach, T. Collins, K. R. McKee, J. Z. Leibo, K. Larson, and T. Graepel. Open Problems in Cooperative AI. *arXiv e-prints*, December 2020.

A. Dafoe, Y. Bachrach, G. Hadfield, E. Horvitz, K. Larson, and T. Graepel. Cooperative ai: machines must learn to find common ground. *Nature*, 7857:33–36, 2021. doi: 10.1038/d41586-021-01170-0.

C. Daskalakis, D.J. Foster, and N. Golowich. Independent Policy Gradient Methods for Competitive Reinforcement Learning. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 5527–5540. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/3b2acfe2e38102074656ed938abf4ac3-Paper.pdf.

Damek Davis and Dmitriy Drusvyatskiy. Stochastic subgradient method converges at the rate $o(k^{-1/4})$ on weakly convex functions. *CoRR*, abs/1802.02988, 2018.

R. Hassin and M. Haviv. *To queue or not to queue: Equilibrium behavior in queueing systems*. Kluwer Academic Publishers, Boston, USA, 2003.

S. Kakade and J. Langford. Approximately Optimal Approximate Reinforcement Learning. In *Proceedings of the Nineteenth International Conference on Machine Learning*, ICML '02, page 267–274, San Francisco, CA, USA, 2002. Morgan Kaufmann Publishers Inc. ISBN 1558608737.

Robert Kleinberg, Georgios Piliouras, and Éva Tardos. Multiplicative updates outperform generic no-regret learning in congestion games. In *ACM Symposium on Theory of Computing (STOC)*, 2009.

E. Koutsoupias and C. Papadimitriou. Worst-case equilibria. In *(STACS)*, pages 404–413. Springer-Verlag, 1999.

Jason D Lee, Ioannis Panageas, Georgios Piliouras, Max Simchowitz, Michael I Jordan, and Benjamin Recht. First-order methods almost always avoid strict saddle points. *Mathematical programming*, 176(1):311–337, 2019.

J. R. Marden. State based potential games. *Automatica*, 48 (12):3075–3088, 2012. ISSN 0005-1098. doi: 10.1016/j.automatica.2012.08.037.

Ruta Mehta, Ioannis Panageas, and Georgios Piliouras. Natural selection as an inhibitor of genetic diversity: Multiplicative weights updates algorithm and a conjecture of haploid genetics. In *Innovations in Theoretical Computer Science*, 2015.

Panayotis Mertikopoulos, Christos Papadimitriou, and Georgios Piliouras. Cycles in adversarial regularized learning. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2703–2717. SIAM, 2018.

D. Mguni. Stochastic Potential Games. *arXiv e-prints*, art. arXiv:2005.13527, May 2020.

D. Mguni, Y. Wu, Y. Du, Y. Yang, Z. Wang, M. Li, Y. Wen, J. Jennings, and J. Wang. Learning in Nonzero-Sum Stochastic Games with Potentials. *arXiv e-prints*, art. arXiv:2103.09284, March 2021.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, Feb 2015. ISSN 1476-4687. doi: 10.1038/nature14236.

D. Monderer and L. S. Shapley. Potential Games. *Games and Economic Behavior*, 14(1):124–143, 1996. ISSN 0899-8256. doi: 10.1006/game.1996.0044.

OpenAI. Openai five. openai.com, 2018.

Ioannis Panageas, Georgios Piliouras, and Xiao Wang. Multiplicative weights update as a distributed constrained optimization algorithm: Convergence to second-order stationary points almost always. In *ICML*, 2018.

T. Roughgarden. Intrinsic robustness of the price of anarchy. *J. ACM*, 62(5), November 2015. ISSN 0004-5411. doi: 10.1145/2806883.

Tim Roughgarden and Éva Tardos. How bad is selfish routing? *Journal of the ACM (JACM)*, 49(2):236–259, 2002.

L. S. Shapley. Stochastic games. *PNAS*, 1953.

David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489, Jan 2016. ISSN 1476-4687. doi: 10.1038/nature16961.

David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, Timothy Lillicrap, Karen Simonyan, and Demis Hassabis. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, 362(6419):1140–1144, 2018. ISSN 0036-8075. doi: 10.1126/science.aar6404.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. A Bradford Book, Cambridge, MA, USA, 2018. ISBN 0262039249.

Vasilis Syrgkanis, Alekh Agarwal, Haipeng Luo, and Robert E. Schapire. Fast convergence of regularized learning in games. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*, NIPS'15, pages 2989–2997, Cambridge, MA, USA, 2015. MIT Press.

S. Valcarcel Macua, J. Zazo, and S. Zazo. Learning Parametric Closed-Loop Policies for Markov Potential Games. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rJm7VfZA-.

Oriol Vinyals, Igor Babuschkin, Wojciech M. Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H. Choi, Richard Powell, Timo Ewalds, Petko Georgiev, Junhyuk Oh, Dan Horgan, Manuel Kroiss, Ivo Danihelka, Aja Huang, Laurent Sifre, Trevor Cai, John P. Agapiou, Max Jaderberg, Alexander S. Vezhnevets, Rémi Leblond, Tobias Pohlen, Valentin Dalibard, David Budden, Yury Sulsky, James Molloy, Tom L. Paine, Caglar Gulcehre, Ziyu Wang, Tobias Pfaff, Yuhuai Wu, Roman Ring, Dani Yogatama, Dario Wünsch, Katrina McKinney, Oliver Smith, Tom Schaul, Timothy Lillicrap, Koray Kavukcuoglu, Demis Hassabis, Chris Apps, and David Silver. Grandmaster level in starcraft ii using multi-agent reinforcement learning. *Nature*, 575 (7782):350–354, Nov 2019. ISSN 1476-4687. doi: 10.1038/s41586-019-1724-z.

Emmanouil-Vasileios Vlatakis-Gkaragkounis, Lampros Flokas, Panayotis Mertikopoulos, and Georgios Piliouras. No-regret learning and mixed nash equilibria: They do not mix. In *Annual Conference on Neural Information Processing Systems*, 2020.

X. Wang and T. Sandholm. Reinforcement Learning to Play an Optimal Nash Equilibrium in Team Markov Games. In *Proceedings of the 15th International Conference on Neural Information Processing Systems*, NIPS'02, page 1603–1610, Cambridge, MA, USA, 2002. MIT Press.

Chen-Yu Wei, Chung-Wei Lee, Mengxiao Zhang, and Haipeng Luo. Last-iterate convergence of decentralized optimistic gradient descent/ascent in infinite-horizon competitive markov games. *CoRR*, abs/2102.04540, 2021.

K. Zhang, Z. Yang, and T. Başar. Multi-Agent Reinforcement Learning: A Selective Overview of Theories and Algorithms. *arXiv e-prints*, art. arXiv:1911.10635, 2019.

Yulai Zhao, Yuandong Tian, Jason D. Lee, and Simon S. Du. Provably efficient policy gradient methods for two-player zero-sum markov games. *CoRR*, abs/2102.08903, 2021.

Global Convergence of Multi-Agent Policy Gradient in Markov Potential Games

Stefanos Leonardos
SUTD

Will Overman
UC Irvine

Ioannis Panageas
UC Irvine

Georgios Piliouras
SUTD

Abstract

Potential games are arguably one of the most important and widely studied classes of normal form games. They define the archetypal setting of multi-agent coordination as all agent utilities are perfectly aligned with each other via a common potential function. Can this intuitive framework be transplanted in the setting of Markov Games? What are the similarities and differences between multi-agent coordination with and without state dependence? We present a novel definition of Markov Potential Games (MPG) that generalizes prior attempts at capturing complex stateful multi-agent coordination. Counter-intuitively, insights from normal-form potential games do not carry over as MPGs can consist of settings where state-games can be zero-sum games. In the opposite direction, Markov games where every state-game is a potential game are not necessarily MPGs. Nevertheless, MPGs showcase standard desirable properties such as the existence of deterministic Nash policies. In our main technical result, we prove (polynomially fast in the approximation error) convergence of independent policy gradient to Nash policies by adapting recent gradient dominance property arguments developed for single agent MDPs to multi-agent learning settings.

1 Introduction

Reinforcement learning (RL) has been a fundamental driver of numerous recent advances in Artificial Intelligence (AI) applications that range from super-human performance in competitive game-playing [30, 31, 5] and strategic decision-making in multiple tasks [23, 25, 35] to robotics, autonomous-driving and cyber-physical systems [7, 39]. A core ingredient for the success of single-agent RL systems, which are typically modelled as Markov Decision Processes (MDPs), is the guarantee of existence of stationary deterministic optimal policies [3, 32]. This allows for the design of efficient algorithms that provably converge towards the optimal policy [1]. However, a majority of the above systems involve multi-agent interactions and despite the notable empirical advancements, there is a lack of understanding about the theoretical convergence guarantees of the existing multi-agent reinforcement learning (MARL) algorithms.

The main challenge when transitioning from single to multi-agent RL settings is the computation of *Nash policies*. A Nash policy for $n > 1$ agents is defined to be a profile of

policies $(\pi_1^*, ..., \pi_n^*)$ so that by fixing the stationary policies of all agents but i, π_i^* is an optimal policy for the resulting single-agent MDP and this is true for all $1 \leq i \leq n$ [1] (see Definition 1). Note that in multi-agent settings, Nash policies *may not be unique* in principle.

A common approach for computing Nash policies in MDPs is the use of *policy gradient* methods. There has been significant progress in the analysis of policy gradient methods during the last couple of years, notably including the works of [1] (and references therein), but it has mainly concerned the single-agent case: the convergence properties of policy gradient in MARL remain poorly understood. Existing steps towards a theory for multi-agent settings involve the papers of [11] who show convergence of *independent policy gradient* to the optimal policy, for two-agent zero-sum stochastic games, of [38] who improve the result of [11] using optimistic policy gradient and of [40] who study extensions of Natural Policy Gradient using function approximation. It is worth noting that the positive results of [11, 38] and [40] depend on the fact that two-agent stochastic zero-sum games satisfy the "min-max equals max-min" property [29] (even though the value-function landscape may not be convex-concave, which implies that Von Neumann's celebrated minimax theorem may not be applicable).

Model and Informal Statement of Results. While the previous works enhance our understanding in *competitive* interactions, i.e., interactions in which gains can only come at the expense of others, MARL in *cooperative* settings remains largely under-explored and constitutes one of the current frontiers in AI research [10, 9]. Based on the above, our work is motivated by the following natural question:

Can we get (provably) fast convergence guarantees for multi-agent RL settings in which cooperation is desirable?

To address this question, we define and study a class of n-agent MDPs that naturally generalize normal form potential games [24], called *Markov Potential Games (MPGs)*. In words, a multi-agent MDP is a MPG as long as there exists a (state-dependent) real-valued potential function Φ so that if an agent i changes their policy (and the rest of the agents keep their policy unchanged), the difference in agent i's value/utility, V^i, is captured by the difference in the value of Φ (see Definition 2). Weighted and ordinal MPGs are defined similar to the normal form counterparts (see Remark 1).

Under our definition, we answer the above motivating question in the affirmative. In particular, we show that if every agent i independently runs (with simultaneous updates) policy gradient on his utility/value V^i, after $O(1/\epsilon^2)$ iterations, the system will reach an ϵ-approximate Nash policy (see informal Theorem 1.1 and formal Theorem 4.5). Moreover, we show the finite sample analogue, that is if every agent i independently runs (with simultaneous updates) stochastic policy gradient, then with high probability, the system will reach an ϵ-approximate Nash policy after $O(1/\epsilon^6)$ iterations.

Along the way, we prove several properties about the structure of MPGs and their Nash policies (see Theorem 1.2 and Section 3). Our results can be summarized in the following two Theorems.

Theorem 1.1 (Convergence of Policy Gradient (Informal)). *Consider a MPG with n agents and let $\epsilon > 0$. Suppose that each agent i runs independent policy gradient using direct*

---

[1] Analogue of Nash equilibrium notion.

parameterization on his policy and that the updates are simultaneous. Then, the learning dynamics reach an ϵ-Nash policy after $\mathcal{O}(1/\epsilon^2)$ iterations. Moreover, suppose that each agent i runs stochastic policy gradient using greedy parameterization (see (4)) on his policy and that the updates are simultaneous. Then the learning dynamics reach an ϵ-Nash policy after $\mathcal{O}(1/\epsilon^6)$ iterations.

This result holds trivially for weighted MPGs and asymptotically also for ordinal MPGs, see Remark 4.

Theorem 1.2 (Structural Properties of MPGs). *The following facts are true for MPGs with n-agents:*

a. *There always exists a Nash policy profile $(\pi_1^*, \ldots, \pi_n^*)$ so that π_i^* is deterministic for each agent i (see Theorem 3.1).*

b. *We can construct MDPs for which each state is an underlying potential game but the MDPs are not MPGs. This can be true regardless of whether the whole MDP is competitive or cooperative in nature (see Examples 1 and 2, respectively). On the opposite side, we can construct MDPs that are MPGs but which include states that are purely competitive (i.e., zero-sum games), see Example 3.*

c. *We provide sufficient conditions so that a MDP is a MPG. These include cases where each state is an underlying potential game and the transition probabilities are not affected by agents actions or the reward functions satisfy certain regularity conditions between different states (see conditions C1 and C2 in Proposition 3.2).*

Technical Overview. The first challenge in the proof of Theorem 1.1 is that multi-agent settings (MPGs) do not satisfy the gradient dominance property, which is an important part in the proof of convergence of policy gradient in single-agent settings [1]. In particular, there is no uniqueness of optimal policies and as a result, there is not a properly defined notion of value in MPGs (in contrast to zero-sum stochastic games [11]). On the positive side, we show that agent-wise (i.e., after fixing the policy of all agents but i), the value function, V^i, satisfies the gradient dominance property along the direction of π_i (policy of agent i). This can be leveraged to show that every *(approximate) stationary point* (Definition 4) of the potential function Φ is an *(approximate) Nash policy* (Lemma 4.2). As a result, convergence to an approximate Nash policy is established by showing that Φ is smooth and then applying *Projected Gradient Ascent* (PGA) on Φ. This step uses the rather well-known fact that (PGA) converges to ϵ-stationary points in $O(1/\epsilon^2)$ iterations for smooth functions. As a result, by applying PGA on the potential Φ, one gets an approximate Nash policy. Our convergence result then follows by showing that PGA on the potential function, Φ, generates the same dynamics as if each agent i runs independent PGA on their value function, V^i.

In the case that agents do not have access to exact gradients, we derive a similar result for finite samples. In this case, we apply *Projected Stochastic Gradient Ascent (PSGA)* on Φ which (as was the case for PGA) can be shown to be the same as when agents apply PSGA independently on their individual value functions. The key is to get an *unbiased sample* for the gradient of the value functions and prove that it has bounded variance (in terms of the parameters of the MPG). This comes from the discount factor, γ; in this case, $1 - \gamma$ can be interpreted as the probability to terminate the MDP at a particular state (and γ to continue).

This can be used to show that a trajectory of the MDP is an unbiased sample for the gradient of the value functions. To guarantee that the estimate has bounded variance, we apply the approach of [11] which requires that agents perform PSGA with α-greedy exploration (see (4)). The main idea is that this parameterization stays away from the boundary of the simplex throughout its trajectory.

Concerning our structural results, the main technical challenge is the dependence of state-transitions (in addition to agents' rewards) on agents' actions. Our work in this part is mainly concerned with showing that the class of MPGs can be significantly larger than state based potential games but also that even simple coordination games may fail to satisfy the (exact) MPG property. Finally, concerning the existence of a deterministic Nash policies, the main challenge is (as in Theorem 1.1) the lack of a value in general multi-agent settings. As we show in the proof of Theorem 3.1, this issue can be still handled within the class of MPGs by constructing single-agent deviations (to deterministic optimal policies) which keep the value of the potential constant (at its global maximum). This process (which leads to a deterministic Nash policy profile) depends critically on the MPG property and does not generalize to arbitrary MARL settings.

Other works on MPGs. There are only a few papers in the recent literature that define and analyze MARL settings under the term *Markov Potential Games* using slight different definitions (see [18, 34]).[2] These papers mainly focus on state-based potential MDPs (i.e., MDPs in which every state is a potential game) and require rather restrictive additional conditions, such as equality or monotonicity of the state-based potential functions, to address the computational challenge of finding Nash policies.[3] Our current results demonstrate the efficiency of simultaneous policy gradient as a to powerful method to find Nash policies even without additional restrictive assumptions on the state-based potential functions. Moreover, as mentioned in Theorem 1.2, the current definition also encompasses MDPs that are not necessarily potential at each state. To the best of our knowledge, the only (cooperative) MPGs that have been successfully addressed prior to this work, are the ones in which all agents receive the same value/utility [37].

2 Preliminaries

Markov Decision Process (MDP). The following notation is standard and largely follows [1] and [11]. We consider a setting with n agents who repeatedly select actions in a shared Markov Decision Process (MDP). The goal of each agent is to maximize their respective value function. Formally, a MDP is defined as a tuple $\mathcal{G} = (\mathcal{S}, \mathcal{N}, \{\mathcal{A}_i, R_i\}_{i \in \mathcal{N}}, P, \gamma, \rho)$, where

- $\mathcal{S}$ is a finite state space of size $S = |\mathcal{S}|$. We will write $\Delta(\mathcal{S})$ to denote the set of all probability distributions over the set $\mathcal{S}$.

- $\mathcal{N} = \{1, 2, \ldots, n\}$ is the set of the $n \geq 2$ agents in the game.

---

[2]Ongoing works also include [21, 22].

[3]The relation of these conditions to the current work is discussed in more detail in Proposition 3.2 and Remark 2.

4

- $\mathcal{A}_i$ is a finite action space for agent $i \in \mathcal{N}$ with generic element $a_i \in \mathcal{A}_i$. Using common conventions, we will write $\mathcal{A} = \prod_{i \in \mathcal{N}} \mathcal{A}_i$ and $\mathcal{A}_{-i} = \prod_{j \neq i} \mathcal{A}_j$ to denote the joint action spaces of all agents and of all agents other than i with generic elements $\mathbf{a} = (a_i)_{i \in \mathcal{N}}$ and $\mathbf{a_{-i}} = (a_j)_{i \neq j \in \mathcal{N}}$, respectively. According to this notation, we have that $\mathbf{a} = (a_i, \mathbf{a_{-i}})$. We will write $X = |\mathcal{X}|$ and $\Delta(\mathcal{X})$ to denote the size of any set $\mathcal{X} \in \{\mathcal{A}_i, \mathcal{A}_{-i}, \mathcal{A}\}$ and the space of all probability distributions over $\mathcal{X}$, respectively.

- $R_i : \mathcal{S} \times \mathcal{A} \to [-1, 1]$ is the individual reward function of agent $i \in \mathcal{N}$, i.e., $R_i(s, a_i, \mathbf{a}_{-i})$ is the instantaneous reward of agent i when agent i takes action a_i and all other agents take actions $\mathbf{a}_{-i}$ at state $s \in \mathcal{S}$.

- P is the transition probability function, for which $P(s' \mid s, \mathbf{a})$ is the probability of transitioning from s to s' when $\mathbf{a} \in \mathcal{A}$ is the action profile chosen by the agents.

- γ is a discount factor for future rewards of the MDP, shared by all agents.

- $\rho \in \Delta(\mathcal{S})$ is the distribution for the initial state at time $t = 0$.

Whenever time is relevant, we will index the above terms with t. In particular, at each time step $t \geq 0$, all agents observe the state $s_t \in \mathcal{S}$, select actions $\mathbf{a}_t = (a_{i,t}, \mathbf{a}_{-i,t})$, receive rewards $r_{i,t} := R_i(s_t, \mathbf{a}_t), i \in \mathcal{N}$ and transition to the next state $s_{t+1} \sim P(\cdot \mid s_t, \mathbf{a}_t)$. We will write $\tau = (s_t, \mathbf{a}_t, \mathbf{r}_t)_{t \geq 0}$ to denote the trajectories of the system, where $\mathbf{r}_t := (r_{i,t}), i \in \mathcal{N}$.

Policies and Value Functions. For each agent $i \in \mathcal{N}$, a deterministic, stationary policy $\pi_i : \mathcal{S} \to \mathcal{A}_i$ specifies the action of agent i at each state $s \in \mathcal{S}$, i.e., $\pi_i(s) = a_i \in \mathcal{A}_i$ for each $s \in \mathcal{S}$. A stochastic, stationary policy $\pi_i : \mathcal{S} \to \Pi_i$, where $\Pi_i := \Delta(\mathcal{A}_i)^S$, specifies a probability distribution over the actions of agent i for each state $s \in \mathcal{S}$. In this case, we will write $a_i \sim \pi_i(\cdot \mid s)$ to denote the randomized action of agent i at state $s \in \mathcal{S}$. As above, we will write $\pi = (\pi_i)_{i \in \mathcal{N}} \in \Pi := \times_{i \in \mathcal{N}} \Delta(\mathcal{A}_i)^S$ and $\pi_{-i} = (\pi_j)_{i \neq j \in \mathcal{N}} \in \Pi_{-i} := \times_{i \neq j \in \mathcal{N}} \Delta(\mathcal{A}_j)^S$ to denote the joint policies of all agents and of all agents other than i, respectively. A joint policy π induces a distribution Pr^π over trajectories $\tau = (s_t, \mathbf{a}_t, \mathbf{r}_t)_{t \geq 0}$, where s_0 is drawn from the initial state distribution ρ and $a_{i,t}$ is drawn from $\pi_i(\cdot \mid s_t)$ for all $i \in \mathcal{N}$.

The value function, $V_s^i : \Pi \to \mathbb{R}$, gives the expected reward of agent $i \in \mathcal{N}$ when $s_0 = s$ and the agents draw their actions, $\mathbf{a}_t = (a_{i,t}, \mathbf{a}_{-i,t})$, at time $t \geq 0$ from policies $\pi = (\pi_i, \pi_{-i})$

$$V_s^i(\pi) := \mathbb{E}_\pi \left[\sum_{t=0}^\infty \gamma^t r_{i,t} \mid s_0 = s \right]. \tag{1}$$

We also denote $V_\rho^i(\pi) = \mathbb{E}_{s \sim \rho}[V_s^i(\pi)]$ if the initial state is random and follows distribution ρ.

Nash Policies. The solution concept that will be focusing on is the Nash Policy. Formally:

Definition 1 (Nash Policy)**.** A joint policy, $\pi^* = (\pi_i^*)_{i \in \mathcal{N}} \in \Pi$, is a Nash policy if for each agent $i \in \mathcal{N}$ it holds that

$$V_s^i(\pi_i^*, \pi_{-i}^*) \geq V_s^i(\pi_i, \pi_{-i}^*), \text{ for all } \pi_i \in \Delta(\mathcal{A}_i)^S, \text{ and all } s \in \mathcal{S},$$

i.e., if the policy, π_i^*, of each agent $i \in \mathcal{N}$ maximizes agent i's value function for each starting state $s \in \mathcal{S}$ given the policies, $\pi_{-i}^* = (\pi_j^*)_{j \neq i}$, of all other agents $j \neq i \in \mathcal{N}$. Similarly, a joint policy $\pi^* = (\pi_i^*)_{i \in \mathcal{N}}$ is an ϵ-Nash policy if there exists an $\epsilon > 0$ so that for each agent i

$$V_s^i(\pi_i^*, \pi_{-i}^*) \geq V_s^i(\pi_i, \pi_{-i}^*) - \epsilon, \text{ for all } \pi_i \in \Delta(\mathcal{A}_i)^S, \text{ and all } s \in \mathcal{S}.$$

We would like to note the definition of Nash policy is the same if $s \sim \rho$ (random starting state).

Markov Potential Games. We are ready to define the class of MDPs that we will focus on for the rest of the paper, i.e., Markov Potential Games.

Definition 2 (Markov Potential Game). A Markov Decision Process (MDP), $\mathcal{G}$, is called a *Markov Potential Game (MPG)* if there exists a (state-dependent) function $\Phi_s : \Pi \to \mathbb{R}$ for $s \in \mathcal{S}$ so that

$$\Phi_s(\pi_i, \pi_{-i}) - \Phi_s(\pi'_i, \pi_{-i}) = V_s^i(\pi_i, \pi_{-i}) - V_s^i(\pi'_i, \pi_{-i}),$$

for all agents $i \in \mathcal{N}$, all states $s \in \mathcal{S}$ and all policies $\pi_i, \pi'_i \in \Pi_i, \pi_{-i} \in \Pi_{-i}$. We should note that by linearity of expectation, it follows that $\Phi_\rho(\pi_i, \pi_{-i}) - \Phi_\rho(\pi'_i, \pi_{-i}) = V_\rho^i(\pi_i, \pi_{-i}) - V_\rho^i(\pi'_i, \pi_{-i})$, where $\Phi_\rho(\pi) := \mathbb{E}_{s \sim \rho} [\Phi_s(\pi)]$.

As in normal-form games, an immediate consequence of this definition is that the value function of each agent in a MPG can be written as a sum of the potential *(common term)* and a term that does not depend on that agent's policy *(dummy term)*, cf. Proposition B.1 in Appendix B, i.e., for each agent $i \in \mathcal{N}$ there exists a function $U_s^i : \Pi_{-i} \to \mathbb{R}$ so that

$$V_s^i(\pi) = \Phi_s(\pi) + U_s^i(\pi_{-i}), \quad \text{for all } \pi \in \Pi.$$

Remark 1 (Ordinal and Weighted Potential Games). Similar to normal-form games, one may also define more general notions of MPGs, such as ordinal or weighted Markov Potential Games. Specifically, if for all agents $i \in \mathcal{N}$, all states $s \in \mathcal{S}$ and all policies $\pi_i, \pi'_i \in \Pi_i, \pi_{-i} \in \Pi_{-i}$, the function $\Phi_s, s \in \mathcal{S}$ satisfies

$$\Phi_s(\pi_i, \pi_{-i}) - \Phi_s(\pi'_i, \pi_{-i}) > 0 \iff V_s^i(\pi_i, \pi_{-i}) - V_s^i(\pi'_i, \pi_{-i}) > 0,$$

then the MDP, $\mathcal{G}$, is called an *Ordinal Markov Potential Game (OMPG)*. If there exist positive constants $w_i > 0, i \in \mathcal{N}$ so that

$$\Phi_s(\pi_i, \pi_{-i}) - \Phi_s(\pi'_i, \pi_{-i}) = w_i(V_s^i(\pi_i, \pi_{-i}) - V_s^i(\pi'_i, \pi_{-i})),$$

then $\mathcal{G}$ is called a *Weighted Markov Potential Game (WMPG)*.

Similarly to normal-form games, such classes are naturally motivated also in the setting of multi-agent MDPs. As Example 2 in Appendix B.1 shows, even simple potential-like settings, i.e., settings in which coordination is desirable for all agents, may fail to be exact MPGs (but may still be ordinal or weighted MPGs) due to the dependence of both the rewards and the transitions on agents' decisions. From our current perspective, ordinal and weighted MPGs (as defined in Remark 1) remain relevant, since as we argue, policy gradient still converges to Nash policies in these classes of games (see Remark 4).

Independent Policy Gradient and Direct Parameterization We assume that all agents update their policies *independently* according to the *projected gradient ascent (PGA)* or *policy gradient* algorithm on their policies. Independence here refers to the fact that (PGA) requires only local information (each agent's own rewards, actions and view of the environment) to

form the updates, i.e., to estimate that agent's policy gradients. Such protocols are naturally motivated and particularly suitable for distributed AI settings in which all information about the interacting agents, the type of interaction and the agent's actions (policies) is encoded in the environment of each agent.[4]

The PGA algorithm is given by

$$\pi_i^{(t+1)} := P_{\Delta(\mathcal{A}_i)^S}\left(\pi_i^{(t)} + \eta \nabla_{\pi_i} V_\rho^i(\pi^{(t)})\right),\tag{PGA}$$

for each agent $i \in \mathcal{N}$, where $P_\Delta(\mathcal{A}_i)^S$ is the projection onto $\Delta(\mathcal{A}_i)^S$ in the Euclidean norm. Here, the additional argument $t \geq 0$ denotes time. We also assume that all players $i \in \mathcal{N}$ use direct policy parameterizations, i.e.,

$$\pi_i(a \mid s) = x_{i,s,a}\tag{2}$$

with $x_{i,s,a} \geq 0$ for all $s \in \mathcal{S}, a \in \mathcal{A}_i$ and $\sum_{a \in \mathcal{A}_i} x_{i,s,a} = 1$ for all $s \in \mathcal{S}$. This parameterization is complete in the sense that any stochastic policy can be represented in this class [1].

In practice, agents use *projected stochastic gradient ascent* (PSGA), according to which, the actual gradient, $\nabla_{\pi_i} V_\rho^i(\pi^{(t)})$, is replaced by an estimate thereof that is calculated from a randomly selected (yet finite) sample of trajectories of the MDP. This estimate, $\hat{\nabla}_{\pi_i}^{(t)}$ may be derived from a single or a batch of observations which in expectation behave as the actual gradient. We choose the estimate of the gradient of V_ρ^i to be

$$\hat{\nabla}_{\pi_i}^{(t)} = R_i^{(T,t)} \sum_{k=0}^{T} \nabla \log \pi_i(a_k^{(t)} \mid s_k^{(t)}),\tag{3}$$

where $s_0^t \sim \rho$, and $R_i^{(T,t)} = \sum_{k=0}^{T} r_{i,t}^k$ is the sum of rewards of agent i for a batch of time horizon T along the trajectory generated by the stochastic gradient ascent algorithm at its t-th iterate (recall that the discount factor, γ, functions as the probability to continue at each step, so T is sampled from a geometric distribution).

The direct parameterization is not sufficient to ensure that the variance of the gradient estimator is bounded (as policies approach the boundary). In this case, we will require that each agent $i \in \mathcal{N}$ uses instead direct parameterization with α-greedy exploration as follows

$$\pi_i(a \mid s) = (1 - \alpha)x_{i,s,a} + \alpha/A_i,\tag{4}$$

where α is the exploration parameter for all agents. Under greedy exploration, it can be shown that (3) is unbiased and has bounded variance for α-greedy exploration (see Lemma 4.6). The form of PSGA is given below:

$$\pi_i^{(t+1)} := P_{\Delta(\mathcal{A}_i)^S}\left(\pi_i^{(t)} + \eta \hat{\nabla}_{\pi_i}^{(t)}\right).\tag{PSGA}$$

---

[4]In practice, even though every agent treats their environment as fixed, the environment changes as other agents update their policies. This is what makes the analysis of such protocols particularly challenging in full general. It also highlights the importance of studying classes of games (MDPs) in which convergence of independent learning protocols can be obtained such as zero-sum stochastic games [11] or MPGs as we do in this paper.

3   Structural Properties of Markov Potential Games

Existence of Deterministic Nash Policies.   The first question that we examine, is whether MPGs possess a *deterministic* optimal policy profile, as is the case in normal-form potential games [24]. In Theorem 3.1, we show that this important property indeed carries over (which settles part (a) of Theorem 1.2).

Theorem 3.1 (Deterministic Optimal Policy Profile)**.** *Let $\mathcal{G}$ be a Markov Potential Game (MPG). Then, there exists a Nash policy $\pi^* \in \Delta(\mathcal{A})^S$ which is deterministic, i.e., for each agent $i \in \mathcal{N}$ and each state $s \in \mathcal{S}$, there exists an action $a_i \in \mathcal{A}_i$ so that $\pi_i^*(a_i \mid s) = 1$.*

The proof of Theorem 3.1 (which is deferred to Appendix B) exploits the fact that we can iteratively reduce the non-deterministic components of an arbitrary Nash policy profile that corresponds to a global maximizer of the potential and still retain the Nash profile property at all times. At each iteration, we isolate an agent $i \in \mathcal{N}$, and find a deterministic (optimal) policy for that agent in the (single-agent) MDP in which the policies of all other agents but i remain fixed. The important observation is that the resulting profile is again a global maximizer of the potential and hence, a Nash policy profile. This argument critically relies on the MPG structure and does not seem directly generalizable to MDPs that do not satisfy Definition 2.

Sufficient Conditions for MPGs.   We next turn to the question of which types of games are captured by Definition 2. It is tempting to think that MDPs which are potential at every state (meaning that the immediate rewards at every state are captured by a (normal-form) potential game at that state) are trivially MPGs. As we show in Examples 1 and 2, this intuition fails in the most straightforward way: we can construct simple MDPs that are potential at every state but which are purely competitive (do not possess a deterministic Nash policy) overall (Example 1) or which are cooperative in nature overall but which do not possess an exact potential function (Example 2).

Example 1. Consider the two-agent, two-state, and two actions per state MDP, $\mathcal{G} = (\mathcal{S} = \{0, 1\}, \mathcal{N} = \{A, B\}, (\mathcal{A}_i = \{0, 1\}, R_i)_{i \in \mathcal{N}}, P, \rho)$ in Figure 1. At state 0 (1), agent A always receives $+2$ (0) and agent B always receives 0 ($+2$) regardless of the actions they choose. That is, the reward functions for both states are constant, which implies that both states are potential games. The transitions are determinstic and are given by

$$s_{t+1} = s_t \oplus a_A^{s_t} \oplus a_B^{s_t},$$

where $\oplus$ denotes the xor operator or equivalently, addition modulo 2, i.e., $1 \oplus 1 = 0$. The MDP $\mathcal{G}$ is illustrated in Figure 1.

To show that $\mathcal{G}$ is not a MPG, it suffices to show that it cannot have a deterministic optimal policy as should be the case according to Theorem 3.1. To obtain a contradiction, assume that agent A is using a deterministic action $a_A^0 \in \{0, 1\}$ at state 0. Then, agent B, who prefers to move to state 1, will optimize their utility by choosing the action $a_B^0 \in \{0, 1\}$ that yields $a_A^0 \oplus a_B^0 = 1$. In other words, given any deterministic action of agent A at state 0, agent B can choose an action that always moves the sequence of play to state 1. Thus, such an action cannot be optimal for agent A which implies that the MDP $\mathcal{G}$ does not have a deterministic optimal policy profile as claimed.

$$a_A^0 \oplus a_B^0 = 0 \qquad\qquad a_A^1 \oplus a_B^1 = 0$$

otherwise

s_0 s_1

otherwise

$$\begin{array}{c} \\ 0 \\ 1 \end{array} \begin{array}{cc} 0 & 1 \\ \left(\begin{array}{cc} 2,0 & 2,0 \\ 2,0 & 2,0 \end{array} \right. \end{array}$$

otherwise

$$\begin{array}{c} \\ 0 \\ 1 \end{array} \begin{array}{cc} 0 & 1 \\ \left(\begin{array}{cc} 0,2 & 0,2 \\ 0,2 & 0,2 \end{array} \right. \end{array}$$
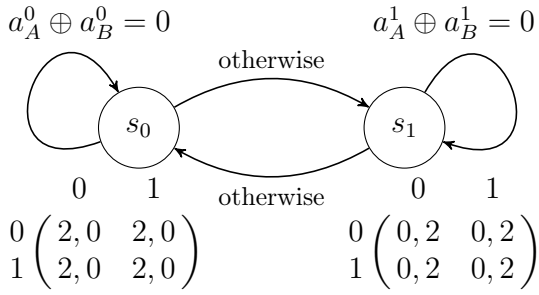
Figure 1: A MDP which is potential at every state but which not a MPG due to conflicting preferences over states. The agents' instantaneous rewards, $(R_A(s, \mathbf{a}), R_B(s, \mathbf{a}))$, are shown in matrix form below each state $s = 0, 1$.

$$(a_A^0, a_B^0) = (0,0)$$

otherwise

s_0 s_1

$(0,0)$

$$\begin{array}{c} \\ 0 \\ 1 \end{array} \begin{array}{cc} 0 & 1 \\ \left(\begin{array}{cc} 5, \ 2 & -1, -2 \\ -5, -4 & 1, \ 4 \end{array} \right. \end{array}$$
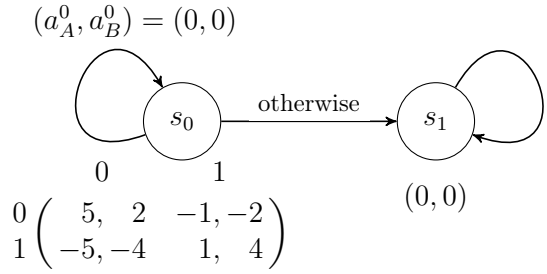
Figure 2: A MDP which is potential at every state and cooperative in nature but which is not a MPG. The action-dependent transitions do not allow the derivation of an exact potential function.

Intuitively, the two agents in Example 1 play a game of *matching pennies* in terms of the actions that they choose (since they prefer playing in opposing states). Thus, competition arises due to the opposing preferences of the agents over states even though the immediate rewards at each states are determined by normal form potential games.

Example 2 shows that a state-based potential game may fail to be a MPG even if agents have similar preferences over states. In that case, the reason is that one cannot find an exact potential function due to the dependence of the transitions on agents' actions. However, in the case of Example 2, it is straightforward to show that the game is an ordinal potential game, cf. Remark 2.

Example 2. Consider the two-agent, two-state MDP, $\mathcal{G} = (\mathcal{S} = \{0, 1\}, \mathcal{N} = \{A, B\}, \{\mathcal{A}_i, R_i\}_{i \in \mathcal{N}}, P, \rho)$ in Figure 2. At state s_0, each agent has two actions, $\mathcal{A}_i = \{0, 1\}$, whereas at state s_1, each agent has a single action. The transitions and instantaneous rewards, $(R_A(s, \mathbf{a}), R_B(s, \mathbf{a})), s = 0, 1, \mathbf{a} = (a_A^s, a_B^s)$ of this MDP are shown in Figure 2. If the action profile $\mathbf{a} = (a_A^0, a_B^0) = (0, 0)$ is selected at state s_0, then the play remains there, otherwise the play transitions to state s_1 and remains there forever.

Since, the game at s_0 is a version of the Battle of the Sexes and hence a potential game (see also Appendix B.1), there exists a potential function ϕ_0, such that we may write the instantaneous reward, $R_i(s_0, \mathbf{a})$ of agent $i = A, B$ at that state as the sum of the potential, $\phi_0(\pi)$ (common term) and a dummy term, $u_0^i(\pi_{-i})$, which does not depend on the action (policy) of agent i, but only on the action (policy) of agent $-i$, i.e., $R_i(s_0, \pi) = \phi_0(\pi) + u_0^i(\pi_{-i})$, for $i = A, B$. Here we are using the slight abuse of notation that $R_i(s, \pi) = \mathbb{E}_{\mathbf{a} \sim \pi} R_i(s, \mathbf{a})$. This leads (after some standard algebra) to the following expression for the value function $V_0^i(\pi)$ of agent $i = A, B$ with starting state s_0

$$V_0^i(\pi) = \frac{\phi_0(\pi)}{1 - \gamma pq} + \frac{u_i(\pi_{-i})}{1 - \gamma pq}, \quad \text{for } i = A, B,$$

where $p, q \in [0, 1]$ are the probabilities with which agents A and B respectively select their action 0 at state s_0. This expression clearly indicates the complexity that emerges in MPGs versus static games. Namely, the first term of the value function is a common term (same for both agents) that can conveniently become part of a potential function. However, the second term is a mixture of a common term (denominator) and a term that is different for each agent

9

(numerator). The reason is that the policy of each agent determines the time that the agents spend at each state and thus, it does not (generally) allow for an agent independent term (as required by the definition of a potential game). However, this game is clearly a *potential-like game* in which agents have common interests. This motivates to look at the notion of ordinal or weighted MPGs. Note that (by a straightforward calculation) this game is an ordinal MPG for the potential function $\Phi_s = \phi_s$ for $s = 0, 1$.

Based on the intuition from the previous Examples, we formulate the following sufficient conditions in Proposition 3.2 which ensure that a state based potential game (i.e., a game that is potential at every state) is also a MPG according to Definition 2 (cf. Theorem 1.2 part (c)).

Proposition 3.2 (Sufficient Conditions for MPGs). *Consider a MDP $\mathcal{G}$ in which every state $s \in \mathcal{S}$ is a potential game, i.e., the immediate rewards $R(s, \mathbf{a}) = (R_i(s, \mathbf{a}))_{i \in \mathcal{N}}$ for each state $s \in \mathcal{S}$ are captured by the utilities of a potential game with potential function ϕ_s. Additionally, assume that one of the following conditions holds*

C1. Agent-Independent Transitions: $P(s' \mid s, \mathbf{a})$ does not depend on $\mathbf{a}$, that is, $P(s' \mid s, \mathbf{a}) = P(s' \mid s)$ is just a function of the present state for all states $s, s' \in \mathcal{S}$.

C2. Equality of Individual Dummy Terms: $P(s' \mid s, \mathbf{a})$ is arbitrary but the dummy terms of each agent's immediate rewards are equal across all states, i.e., there exists a function $u^i : \Delta(\mathcal{A}_{-i})^S \to \mathbb{R}$ such that

$$R_i(s, a_i, \mathbf{a}_{-i}) = \phi_s(\pi_i, \pi_{-i}) + u^i(\pi_{-i}), \text{ for all states } s \in \mathcal{S}.$$

If either C1 or C2 are true, then $\mathcal{G}$ is a MPG.

Remark 2. As the rest of the proofs of Section 3, the proof of Proposition 3.2 is provided in Appendix B. The following remarks are due.

1. Condition C2 (or variations of it) is also known as *state-transitivity* and is present as requirement in the existing definitions of potential-like MDPs, see e.g., [18, 21, 22] and along with some additional conditions on the transitions also in [34]. Example 2 shows that such conditions are restrictive, in the sense that they do not capture very simple MDPs that intuitively have a potential-like (cooperative) structure. These observations motivate the study ordinal or weighted potential games as natural models to capture instances that are of cooperative nature but which are not MPGs (cf. Remark 1). As we show, our convergence results about independent policy gradient naturally extend to these classes as well (see Remark 4).

2. Another condition that is (trivially) sufficient to ensure that a MDP that is potential at every state is also a MPG is that the instantaneous rewards of all agents are the same at each state, i.e., that $R_i(s, a_i, \mathbf{a}_{-i}) = \phi_s(a_i, \mathbf{a}_{-i})$ for all agents $i \in \mathcal{N}$, all actions $a_i \in \mathcal{A}_i$ and all states $s \in \mathcal{S}$. MDPs that satisfy this condition are called *Team Markov Games* and their study (in terms of learning their equilibria) boils down to a single agent setting [37].

The previous discussion focuses on games that are potential at every state as natural candidates to generalize the notion of normal-form games to state games. This leaves an important question unanswered: are there games which are not potential at every state but which are captured by the our current definition of MPGs? Example 3 answers this question affirmatively. Together with Example 1, this settles the claim in Theorem 1.2, part (b).
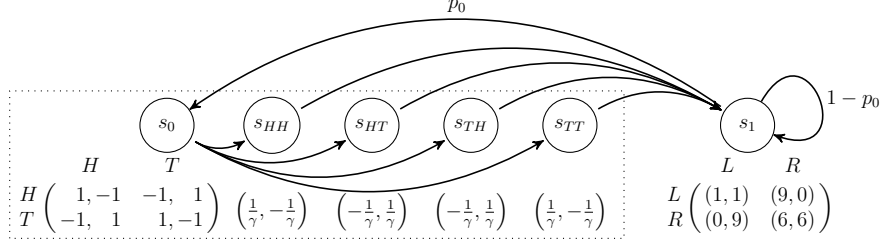
Figure 3: A 2-player MPG which is not potential at every state. The rewards in state s_1 form a potential game, whereas the rewards in s_0 do not. However, the states inside the dotted rectangle do form a potential game and this can be leveraged to show that the whole MPG is a potential game whenever p_0 does not depend on agents' actions.

Example 3 (Not potential at every state may still imply MPG)**.** Consider the 2-agent MDP of Figure 3. At state s_0, agents' rewards, $(R_1(s_0, \mathbf{a}), R_2(s_0, \mathbf{a}))$ form a constant sum (equivalent to zero-sum) game. The agents' actions at s_0 induce a deterministic transition to a state s_{ab} with $a, b \in \{H, T\}$ in which the only available actions are precisely the chosen actions at s_0. Each agent's instantaneous reward at this state is the reward of the other agent at s_0 (scaled by $1/\gamma$). The MDP then transitions deterministically to state s_1 which is a potential game with rewards $(R_1(s_1, \mathbf{a}), R_2(s_1, \mathbf{a}))$. After the agents select their actions at s_1, there is an exogenous given probability, p_0, according to which the play transitions to state $s = 0$. Otherwise it remains at s_1.

While the game at state $s = 0$ is not a potential game, the combined states in the dotted rectangle of Figure 3 do form a potential game, with potential function equal to the sum of the agent's payoffs at s_0 (the rewards of both agents are equal for every pass of the play through the states in the dotted rectangle). Thus, it is not hard to see that both value functions are of the form

$$V_s^i(\pi_1, \pi_2) = c_1(s) \cdot \mathbf{x}_0 \left(R_1(s_0, \mathbf{a}) + R_2(s_0, \mathbf{a}) \right) \mathbf{y}_0 + c_2(s) \cdot \mathbf{x}_1 R_i(s_1, \mathbf{a}) \mathbf{y}_1,$$

for $s \in \{s_0, s_1\}$ and $i = \{1, 2\}$, where $c_1(s), c_2(s) > 0$ are appropriate constants that depend only the state $s \in \{s_0, s_1\}$ and not on the agents. Since the game at s_1 is a potential game, with potential function given by a 2×2 matrix ϕ_1, it is immediate to see that

$$\Phi_s(\pi_1, \pi_2) := c_1(s) \mathbf{x}_0 \left(R_1(s_0, \mathbf{a}) + R_2(s_0, \mathbf{a}) \right) \mathbf{y}_0 + c_2(s) \cdot \mathbf{x}_1 \phi_1 \mathbf{y}_1, \text{ for } s \in \{s_0, s_1\}$$

is a potential function so that $\mathcal{G}$ satisfies the definition of an MPG.

4 Convergence of Policy Gradient in Markov Potential Games

The current section presents the proof of convergence of (projected) policy gradient to approximate Nash policies in Markov Potential Games (MPGs). We analyze the cases of both infinite and finite samples using direct and α-greedy parameterizations, respectively.

Before we proceed with the formal statements and proofs of this section, we provide the definition of distribution mismatch coefficient as it is commonly used [14] applied to our setting.

Definition 3 (Distribution Mismatch coefficient). Let μ be any distribution in $\Delta(\mathcal{S})$ and let $\mathcal{O}$ be the set of policies $\pi \in \Delta(\mathcal{A})^S$ so that π is 0.1-Nash policy. We call

$$D := \max_{\pi \in \mathcal{O}} \left\| \frac{d^\pi_{s_0}}{\mu} \right\|_\infty$$

the *distribution mismatch coefficient*, where d^π_s is the discounted state distribution (17).

The first auxiliary Lemma has to do with the projection operator that is used on top of the independent policy gradient, so that the policy vector $\pi^{(t)}_i$ remains a probability distribution for all agents $i \in \mathcal{N}$ (see (PGA)). It is not hard to show (due to separability) that the projection operator being applied independently for each agent i on $\Delta(\mathcal{A}_i)^S$ is the same as jointly applying projection on $\Delta(\mathcal{A})^S$. This fact is captured in lemma 4.1.

Lemma 4.1 (Projection Operator). *Let $\pi := (\pi_1, ..., \pi_n)$ be the policy profile for all agents and let*

$$\pi' = \pi + \eta \nabla_\pi \Phi_\rho(\pi),$$

be a gradient step on the potential function for a step-size $\alpha > 0$. Then, it holds that

$$P_{\Delta(\mathcal{A})^S}(\pi') = (P_{\Delta(\mathcal{A}_1)^S}(\pi'_1), \ldots, P_{\Delta(\mathcal{A}_n)^S}(\pi'_n)).$$

The main implication of Lemma 4.1 along with the equality of the derivatives between value functions and the potential function in MPGs, i.e., $\nabla_{\pi_i} V^i_s(\pi) = \nabla_{\pi_i} \Phi(\pi)$ for all $i \in \mathcal{N}$ (see property P2 in Proposition B.1), is that running independent (PGA) on each agent's value function is equivalent to running (PGA) on the potential function Φ. In turn, Lemma 4.2 suggests that as long as policy gradient reaches a point $\pi^{(t)}$ with small gradient along the directions in $\Delta(\mathcal{A})^S$, it must be the case that $\pi^{(t)}$ is an approximate Nash policy. Together with Lemma 4.1, this will be sufficient to prove convergence of (PGA).

Stationary point of Φ. The formal definition of a stationary point for the potential function Φ is given below.

Definition 4 (ϵ-stationary). A policy $\pi := (\pi_1, ..., \pi_n) \in \Delta(\mathcal{A})^S$ is called ϵ-stationary for Φ w.r.t distribution μ as long as

$$\max_{(\pi_1 + \delta_1, ..., \pi_n + \delta_n) \in \Delta(\mathcal{A})^S, \sum_{i \in \mathcal{N}} \|\delta_i\|_2^2 \leq 1} \sum_{i \in \mathcal{N}} \delta_i^\top \nabla_{\pi_i} \Phi_\mu(\pi) \leq \epsilon \tag{5}$$

In words, the function $\Phi(\pi)$ cannot increase in value by more than ϵ along every possible local direction δ that is feasible (namely $\pi + \delta$ is also a policy).

Lemma 4.2 (Stationarity of Φ implies Nash). *Let $\epsilon \geq 0$, and let π be an ϵ-stationary point of Φ (see Definition 4). Then, it holds that π is a $\frac{\sqrt{S}D\epsilon}{1-\gamma}$-Nash policy.*

To prove Lemma 4.2, we will need the Gradient Domination property that has been shown to hold in single-agent MDPs [1]. This is presented in Lemma 4.3.

Lemma 4.3 (Agent-wise Gradient Domination Property in MPGs [1]). *Let $\mathcal{G}$ be a MPG with potential function Φ, fix any agent $i \in \mathcal{N}$, and let $\pi = (\pi_i, \pi_{-i}) \in \Delta(\mathcal{A})^S$ be a policy. Let π_i^* be an optimal policy for agent i in the single agent MDP in which the rest of the agents are fixed to choose π_{-i}. Then, for the policy $\pi^* = (\pi_i^*, \pi_{-i}) \in \Delta(\mathcal{A})^S$ that differs from π only in the policy component of agent i, it holds that*

$$\Phi_\rho(\pi^*) - \Phi_\rho(\pi) \leq \frac{1}{1-\gamma} \left\| \frac{d_\rho^{\pi^*}}{\mu} \right\|_\infty \max_{\pi'=(\pi_i',\pi_{-i})} (\pi' - \pi)^\top \nabla_{\pi_i} \Phi_\mu(\pi),$$

for any distributions $\mu, \rho \in \Delta(\mathcal{S})$.

Remark 3 (Best Response). Intuitively, Lemma 4.3 implies that there is a *best response* structure in the agents' updates that we can exploit to show convergence of (projected) policy gradient to an optimal policy profile. In particular, given a fixed policy profile of all agents other than i, the decision of agent i is equivalent to the decision of that agent in a single MDP. Thus, the inequality (which stems directly from the gradient domination property in the single MDP)

$$V_s^i(\pi^*) - V_s^i(\pi) \leq \frac{1}{1-\gamma} \left\| \frac{d_s^{\pi^*}}{\mu} \right\|_\infty \max_{\pi'=(\pi_i',\pi_{-i})} (\pi' - \pi)^\top \nabla_{\pi_i} V_\mu^i(\pi)$$

implies that any stationary point of V_s^i (w.r.t the variables $x_{i,s,a}$ of agent's i policy with the rest of the variables being fixed) is an optimal policy for i, i.e., a best response given the policies of all other agents.

Lemma 4.3 also suggests that there is an important difference in the Gradient Domination Property between (multi-agent) MPGs and single agent MDPs (cf. Lemma 4.1 in [1]). Specifically, for MPGs, the value (of each agent) at different Nash policies may not be unique[5] which implies that the gradient domination property, as stated in Lemma 4.3, will only be enough to guarantee convergence to *one of* the optimal (stationary) points of Φ (and not necessarily to the absolute maximum of Φ). Having all these in mind, we can now prove Lemma 4.2.

Proof of Lemma 4.2. Fix agent i and suppose that i deviates to an optimal policy π_i^* (w.r.t the corresponding single agent MDP). Since π is ϵ-stationary it holds that (Definition 4)

$$\max_{\pi_i' \in \Delta(\mathcal{A}_i)^S} (\pi_i' - \pi_i)^\top \nabla_{\pi_i} \Phi_\mu(\pi) \leq \sqrt{S}\epsilon. \tag{6}$$

Thus, with $\pi^* = (\pi_i^*, \pi_{-i})$, Lemma 4.3 implies that

$$\begin{aligned}
\Phi_\rho(\pi^*) - \Phi_\rho(\pi) &\leq \frac{1}{1-\gamma} \left\| \frac{d_\rho^{\pi^*}}{\mu} \right\|_\infty \max_{\pi'=(\pi_i',\pi_{-i})} (\pi' - \pi)^\top \nabla_{\pi_i} \Phi_\mu(\pi) \\
&\stackrel{(6)}{\leq} \frac{D}{1-\gamma} \sqrt{S}\epsilon.
\end{aligned} \tag{7}$$

Thus, using the definition of the potential function (cf. Definition 2), we obtain that

$$V_\rho^i(\pi^*) - V_\rho^i(\pi) = \Phi_\rho(\pi^*) - \Phi_\rho(\pi) \leq \frac{\sqrt{S}D\epsilon}{1-\gamma}.$$

---

[5]This is in contrast to single-agent MDPs for which the agent has a unique optimal value despite that the agent's optimal policy is not necessarily unique.

Since the choice of i was arbitrary, we conclude that π is an $\frac{\sqrt{S}D\epsilon}{1-\gamma}$-approximate Nash policy. $\quad\square$

The last critical step before we proceed to the formal statement and proof of Theorem 1.1 is that the potential function Φ is smooth. This fact is used in the analysis of both (PGA) and its stochastic counterpart (PSGA).

Lemma 4.4 (Smoothness of Φ). *Let $A_{\max} := \max_{i \in \mathcal{N}} |\mathcal{A}_i|$ (the maximum number of actions for some agent). Then, for any initial state $s_0 \in \mathcal{S}$ (and hence for every distribution $\mu \in \Delta(\mathcal{S})$ on states) it holds that*

$$\|\nabla_\pi \Phi_{s_0}(\pi) - \nabla_\pi \Phi_{s_0}(\pi')\|_2 \leq \frac{2n\gamma A_{\max}}{(1-\gamma)^3} \|\pi - \pi'\|_2 \tag{8}$$

i.e., $\Phi_\mu(\pi)$ is $\frac{2n\gamma A_{\max}}{(1-\gamma)^3}$-smooth.

Exact gradients case. We are now ready to prove Theorem 1.1 (restated formally), following standard arguments about (PGA). Recall that $A_{\max} := \max_{i \in \mathcal{N}} |\mathcal{A}_i|$ is the maximum number of actions for some agent (which scales *linearly* in the number of agents). Note that the global maximum among all values/utilities of agents must be at most one.

Theorem 4.5 (Formal Theorem 1.1, part (a)). *Let $\mathcal{G}$ be a MPG and let $s_0 \in \mathcal{S}$ denote an arbitrary initial state. Let also $A_{\max} = \max_i |\mathcal{A}_i|$, and set the number of iterations to be $T = \frac{16n\gamma D^2 S A_{\max}}{(1-\gamma)^5 \epsilon^2}$ and the learning rate (step-size) to be $\eta = \frac{(1-\gamma)^3}{2n\gamma A_{\max}}$. If the agents run independent projected policy gradient (PGA) starting from arbitrarily initialized policies, then there exists a $t \in \{1, \ldots, T\}$ such that $\pi^{(t)}$ is an ϵ-approximate Nash policy.*

Proof. The first step is to show that Φ is a β-smooth function, in particular, that $\nabla_\pi \Phi$ is β-Lipschitz with $\beta = \frac{2n\gamma A_{\max}}{(1-\gamma)^3}$ as established in Lemma 4.4. Then, a standard *ascent lemma* for Gradient Ascent (see Lemma D.1 from [6]) implies that for any β-smooth function f it holds that $f(x') - f(x) \geq \frac{1}{2\beta} \|x' - x\|_2^2$ where x' is the next iterate of (PGA). Applied to our setting, this gives

$$\Phi_\mu(\pi^{(t+1)}) - \Phi_\mu(\pi^{(t)}) \geq \frac{(1-\gamma)^3}{4n\gamma A_{\max}} \left\|\pi^{(t+1)} - \pi^{(t)}\right\|_2^2 \tag{9}$$

Thus, if the number of iterates, T, is $\frac{16n\gamma D^2 S A_{\max}}{(1-\gamma)^5 \epsilon^2}$, then there must exist a $1 \leq t \leq T$ so that $\left\|\pi^{(t+1)} - \pi^{(t)}\right\|_2 \leq \frac{\epsilon(1-\gamma)}{2D\sqrt{S}}$. Using a standard approximation property (see Lemma D.2), we then conclude that $\pi^{(t+1)}$ will be a $\frac{\epsilon(1-\gamma)}{D\sqrt{S}}$-stationary point for the potential function Φ. Hence, by Lemma 4.2, it follows that $\pi^{(t+1)}$ is an ϵ-Nash policy and the proof is complete.

Finite sample case In the case of finite samples, we analyze (PSGA) on the value V^i of each agent i which (as was the case for PGA) can be shown to be the same as applying projected gradient ascent on Φ. The key is to get an estimate of the gradient of Φ (3) at every iterate. Note that $1 - \gamma$ now captures the probability for the MDP to terminate after each round (and it does not play the role of a discounted factor since we consider finite length trajectories). Lemma 4.6 argues that the estimator of equation (3) is both unbiased and bounded.

14

Lemma 4.6 (Unbiased estimator with bounded variance). *It holds that $\hat{\nabla}_{\pi_i}^{(t)}$ is an unbiased estimator of $\nabla_{\pi_i}\Phi$ for all $i \in \mathcal{N}$, that is*

$$\mathbb{E}_{\pi^{(t)}}\hat{\nabla}_{\pi_i}^{(t)} = \nabla_{\pi_i}\Phi_\mu(\pi^{(t)}) \text{ for all } i \in \mathcal{N}.$$

Moreover, for all agents $i \in \mathcal{N}$, it holds that

$$\mathbb{E}_{\pi^{(t)}}\left\|\hat{\nabla}_{\pi_i}^{(t)}\right\|_2^2 \leq \frac{24A_{\max}^2}{\epsilon(1-\gamma)^4}, \text{ for all } i \in \mathcal{N}.$$

Proof. It is straightforward from Lemma D.4 and the equality of the partial derivatives between the value functions and the potential, i.e., $\nabla_{\pi_i}\Phi_\mu = \nabla_{\pi_i}V_\mu^i$ for all $i \in \mathcal{N}$ (see property P2 in Proposition B.1). $\square$

We now state and prove part (b) of Theorem 1.1.

Theorem 4.7 (Formal Theorem 1.1, part (b)). *Let $\mathcal{G}$ be a MPG and let $s_0 \in \mathcal{S}$ denote an arbitrary initial state. Let $A_{\max} = \max_i |\mathcal{A}_i|$, and set the number of iterations to be $T = \frac{48(1-\gamma)A_{\max}D^4S^2}{\epsilon^6\gamma^3}$ and the learning rate (step-size) to be $\eta = \frac{\epsilon^4(1-\gamma)^3\gamma}{48nD^2A_{\max}^2S}$. If the agents run projected stochastic policy gradient (PSGA) starting from arbitrarily initialized policies and using α-greedy parametrization with $\alpha = \epsilon^2$, then there exists a $t \in \{1,\ldots,T\}$ such that in expectation, $\pi^{(t)}$ is an ϵ-approximate Nash policy.*

Proof. Let $\delta_t = \hat{\nabla}_\pi^{(t)} - \nabla_\pi\Phi_\mu(\pi^{(t)})$ and set $\lambda = \frac{(1-\gamma)^3}{2n\gamma A_{\max}}$ (the inverse of the smooth parameter in 4.4). Moreover, we set $y^{(t+1)} = P_{\Delta(\mathcal{A})^S}(\pi^{(t)} + \eta\nabla_\pi\Phi(\pi^{(t)}))$ ($y^{(t+1)}$ captures the next iterate of the projected (deterministic) gradient ascent).

We follow the analysis of Projected Stochastic Gradient Ascent for non-convex smooth-functions (see [12], Theorem 2.1) that makes use of the Moreau envelope. Let

$$\phi_\lambda(x) = \underset{y \in \Delta(\mathcal{A})^S}{\arg\min}\left\{-\Phi_\mu(y) + \frac{1}{\lambda}\|x - y\|_2^2\right\},$$

(definition of Moreau envelope for our objective Φ). From the definition of ϕ and a standard property of projection we get

$$\begin{aligned}
\phi_\lambda(\pi^{(t+1)}) &\leq -\Phi_\mu(y^{(t+1)}) + \frac{1}{\lambda}\left\|\pi^{(t+1)} - y^{(t+1)}\right\|_2^2 \\
&\leq -\Phi_\mu(y^{(t+1)}) + \frac{1}{\lambda}\left\|\pi^{(t)} + \eta\hat{\nabla}_\pi^{(t)} - y^{(t+1)}\right\|_2^2 \\
&= -\Phi_\mu(y^{(t+1)}) + \frac{1}{\lambda}\left\|\pi^{(t)} - y^{(t+1)}\right\|_2^2 + \frac{\eta^2}{\lambda}\left\|\hat{\nabla}_\pi^{(t)}\right\|_2^2 + \frac{2\eta}{\lambda}(\pi^{(t)} - y^{(t+1)})^\top\hat{\nabla}_\pi^{(t)}
\end{aligned} \quad (10)$$

Since $\hat{\nabla}_\pi^{(t)}$ is unbiased (Lemma 4.6) we have that $\mathbb{E}[\delta_t|\pi^{(t)}] = 0$, therefore $\mathbb{E}\left[\delta_t^\top(y^{(t+1)} - \pi^{(t)})\right] = 0$. Additionally, by Lemma 4.6 (applied for all agents i) we also have $\mathbb{E}\left[\left\|\hat{\nabla}_\pi^{(t)}\right\|_2^2\right] \leq \frac{24nA_{\max}^2}{\epsilon(1-\gamma)^4}$. Hence by taking expectation on (10) we have:

$$\mathbb{E}[\phi_\lambda(\pi^{(t+1)})] \leq \mathbb{E}\left[-\Phi_\mu(y^{(t+1)}) + \frac{1}{\lambda}\left\|\pi^{(t)} - y^{(t+1)}\right\|_2^2\right] + \frac{2\eta}{\lambda}\mathbb{E}[(\pi^{(t)} - y^{(t+1)})^\top\nabla_\pi\Phi_\mu(\pi^{(t)})] + \frac{24\eta^2nA_{\max}^2}{\lambda\epsilon(1-\gamma)^4}.$$

15

Using the definition of Moreau envelope and the fact that Φ is $\frac{1}{\lambda}$-smooth (Lemma 4.4, after the parametrization, the smoothness parameter does not increase) we conclude that

$$\mathbb{E}[\phi_\lambda(\pi^{(t+1)})] \leq \mathbb{E}[\phi_\lambda(\pi^{(t)})] + \frac{2\eta}{\lambda}\mathbb{E}[(\pi^{(t)} - y^{(t+1)})^\top \nabla_\pi \Phi_\mu(\pi^{(t)})] + \frac{24\eta^2 n A_{\max}^2}{\lambda\epsilon(1-\gamma)^4}$$

$$\leq \mathbb{E}[\phi_\lambda(\pi^{(t)})] + \frac{2\eta}{\lambda}\mathbb{E}\left[\Phi_\mu(\pi^{(t)}) - \Phi_\mu(y^{(t+1)}) + \frac{1}{2\lambda}\left\|\pi^{(t)} - y^{(t+1)}\right\|_2^2\right] + \frac{24\eta^2 n A_{\max}^2}{\lambda\epsilon(1-\gamma)^4},$$

or equivalently

$$\mathbb{E}[\phi_\lambda(\pi^{(t+1)})] - \mathbb{E}[\phi_\lambda(\pi^{(t)})] \leq \frac{2\eta}{\lambda}\mathbb{E}\left[\Phi_\mu(\pi^{(t)}) - \Phi_\mu(y^{(t+1)}) + \frac{1}{2\lambda}\left\|\pi^{(t)} - y^{(t+1)}\right\|_2^2\right] + \frac{24\eta^2 n A_{\max}^2}{\lambda\epsilon(1-\gamma)^4} \tag{11}$$

Adding telescopically (11), dividing by T and because w.l.o.g $-\Phi \in [-1, 0]$, we get that

$$\frac{1}{T} + \frac{24\eta^2 n A_{\max}^2}{\lambda\epsilon(1-\gamma)^4} \geq \frac{2\eta}{\lambda T}\sum_{t=1}^T \mathbb{E}\left[\Phi_\mu(y^{(t+1)}) - \Phi_\mu(\pi^{(t)})\right] - \frac{\eta}{\lambda^2 T}\sum_{t=1}^T \mathbb{E}\left[\left\|y^{(t+1)} - \pi^{(t)}\right\|_2^2\right]$$

$$\geq \min_{t\in[T]}\left\{\frac{2\eta}{\lambda}\mathbb{E}\left[\Phi_\mu(y^{(t+1)}) - \Phi_\mu(\pi^{(t)})\right] - \frac{\eta}{\lambda^2}\mathbb{E}[\left\|y^{(t+1)} - \pi^{(t)}\right\|_2^2]\right\} \tag{12}$$

Let $t*$ be the time index that minimizes the above. We show the following inequality (which provides a lower bound on the RHS of (12):

$$\mathbb{E}\left[\Phi_\mu(y^{(t*+1)}) - \Phi_\mu(\pi^{(t*)})\right] - \frac{1}{2\lambda}\mathbb{E}\left[\left\|y^{(t*+1)} - \pi^{(t*)}\right\|_2^2\right] \geq \frac{1}{\lambda}\mathbb{E}\left[\left\|y^{(t*+1)} - \pi^{(t*)}\right\|_2^2\right] \tag{13}$$

Observe that by $\frac{1}{\lambda}$-smoothness of Φ_μ we get that $H(x) := -\Phi_\mu(x) + \frac{1}{\lambda}\left\|x - \pi^{(t*)}\right\|_2^2$ is $\frac{1}{\lambda}$-strong convex and moreover $y^{(t+1)}$ is the minimizer of H. Therefore we get that

$$H(\pi^{t*}) - H(y^{t*+1}) \geq \frac{1}{2\lambda}\left\|\pi^{t*} - y^{t*+1}\right\|_2^2,$$

or equivalently

$$\Phi_\mu(y^{t*+1}) - \Phi_\mu(\pi^{t*}) - \frac{1}{\lambda}\left\|\pi^{t*} - y^{t*+1}\right\|_2^2 \geq \frac{1}{2\lambda}\left\|\pi^{t*} - y^{t*+1}\right\|_2^2.$$

By taking expectation of terms above, (13) follows. Combining (12) with (13) we conclude that

$$\frac{1}{T} + \frac{24\eta^2 n A_{\max}^2}{\lambda\epsilon(1-\gamma)^4} \geq \frac{2\eta}{\lambda^2}\mathbb{E}\left[\left\|\pi^{t*} - y^{t*+1}\right\|_2^2\right].$$

By Jensen's inequality it occurs that

$$\mathbb{E}\left[\left\|y^{(t*+1)} - \pi^{(t*)}\right\|_2\right] \leq \sqrt{\frac{\lambda^2}{2\eta T} + \eta\frac{12 n\lambda A_{\max}^2}{\epsilon(1-\gamma)^4}}. \tag{14}$$

16

To get an ϵ-Nash policy, we have to bound $\left\|y^{(t*+1)} - \pi^{(t*)}\right\|_2 \leq \frac{\epsilon(1-\gamma)}{2D\sqrt{S}(2+\sqrt{nA_{max}})}$ and choose $\alpha = \epsilon^2$ in the greedy parametrization. This is true because of Lemma D.3 Lemma 4.2. Hence, we need to choose η, T so that

$$\sqrt{\frac{\lambda^2}{2\eta T} + \eta \frac{12n\lambda A_{\max}^2}{\epsilon^2(1-\gamma)^4}} \leq \frac{\epsilon(1-\gamma)}{2D\sqrt{S}(2+\sqrt{nA_{max}})}.$$

We conclude that η can be chosen to be $\frac{\epsilon^4(1-\gamma)^3\gamma}{48nD^2A_{\max}^2S}$ and T to be $\frac{48(1-\gamma)A_{\max}D^4S^2}{\epsilon^6\gamma^3}$. $\qquad\square$

We conclude this section by giving a remark on Weighted and Ordinal MPGs (cf. Definition in 1).

Remark 4 (Weighted and ordinal MPGs). It is rather straightforward to see that our results carry over for weighted MPGs. The only difference in the running time of (PGA) is to account for the weights (which are just multiplicative constants).

In contrast, the extension to ordinal MPGs is not immediate and the reason is that we cannot prove any bound on the smoothness of Φ in that case (i.e., we cannot generalize Lemma 4.4). Therefore, we cannot have rates of convergence of policy gradient. Nevertheless, it is quite straightforward that (PGA) converges asymptotically to critical points (in bounded domains) for differentiable functions. Therefore as long as Φ is differentiable, it is guaranteed that asymptotically (PGA) will converge to a critical point of Φ. By Lemma 4.2, this point will be a Nash policy.

$\qquad\square$

5   Experiments: Congestion Games

We next study the performance of the policy gradient algorithm in a general class of MPGs that are congestion games at every state.

Experimental setup: We consider a MDP in which every state is a congestion game (cf. [4]). In the current experiment, there are $N = 8$ agents, $A_i = 4$ facilities (resources or locations) that the agents can select from and $S = 2$ states: a *safe* state and a *distancing* state. In both states, all agents prefer to be in the same facility with as many other agents as possible *(follow the crowd)* [13]. In particular, the reward of each agent for being at facility k is equal to a predefined positive weight w_k^{safe} times the number of agents at $k = A, B, C, D$. The weights satisfy $w_A^{\text{safe}} < w_B^{\text{safe}} < w_C^{\text{safe}} < w_D^{\text{safe}}$, i.e., facility D is the most preferable by all agents. However, if more than $4 = N/2$ agents find themselves in the same facility, then the game transitions to the distancing state. At the distancing state, the reward structure is the same for all agents, but the weights are reduced by a constant factor, i.e., $w_k^{\text{dist}} = w_k^{\text{safe}} - c$, where $c > 0$ is a (considerably large) constant for $k = A, B, C, D$. To return to the safe state, the agents need to achieve maximum distribution over the facilities, i.e., no more than $2 = N/4$ agents may be in the same facility. We consider deterministic transitions, however, the results are quantitatively equivalent also when these transitions occur with some probability (see Appendix E).
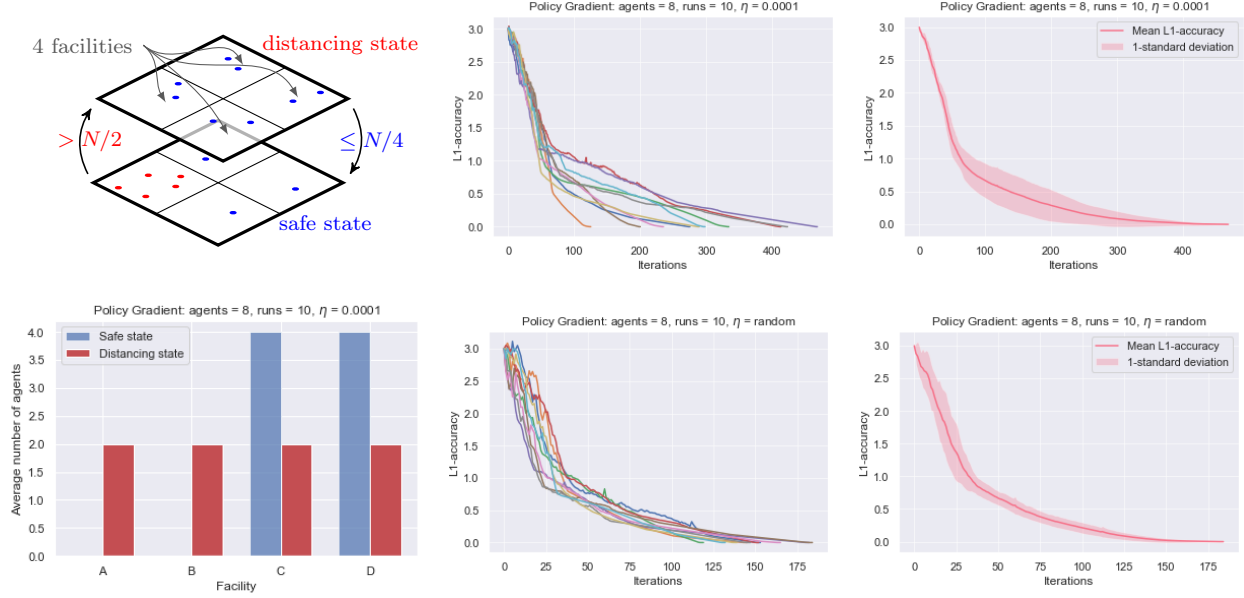
17

Figure 4: Upper left panel: An illustration of the MPG that is used in the experiments with $S = 2$ states, 4 facilities and $N = 8$ agents (description in text). Lower left panel: The distribution of agents at the equilibrium that is reached by the policy gradient algorithm (common for all runs). Right column: Trajectories of the L1-accuracy (average difference between current policy and Nash policy) over the 10 runs for both equal (upper panels) and different learning rates among agents (lower panels).

To see that this MDP is a MPG, it suffices to check that every state is a potential game and that condition C2 (i.e., equality of individual dummy terms) of Proposition 3.2 is satisfied. The first claim is straightforward since at each state, the agents play a congestion game [24, 27]. The second claim follows from the fact that the rewards of all agents in all facilities at the distancing state are shifted by the same constant amount, c. (If c is different for each facility, then the resulting MDP is not an MPG but cooperation is still desirable. For this case, see Appendix E). The MPG is illustrated in the upper left panel of Figure 4.

Paremeters: We perform episodic updates with $T = 20$ steps. At each iteration, we estimate the Q-functions, the value function, the discounted visitation distributions and, hence, the policy gradients using the average of mini-batches of size 20. We use $\gamma = 0.99$. For the presented plots, we use a common learning rate $\eta = 0.0001$ (upper panels) or randomly generated learning rates (different for each agent) in $[.00005, .0005]$ (lower panels) in Figure 4. Note that these learning rates are (several orders of magnitude) larger than the theoretical guarantee, $\eta = \frac{(1-\gamma)^3}{2\gamma A_{\max} n} \approx 1e - 08$, of Theorem 4.5. Experiments (not presented here) with even larger learning rates (e.g., $\eta = 0.001$) did not lead (consistently) to convergence.

Results: The lower left panel of Figure 4 shows that the agents learn the expected Nash profile in both states in all runs (this is common for both the fixed and the random learning rates). At the safe state, the agents distribute themselves equally among the two most preferable facilities (C and D). This leads to a maximum utility and avoids a transition to the distancing

(bad) state. At the distancing state, the agents learn the unique distribution (2 agents per facility) that leads them back to the safe state. Importantly, this (Nash) policy profile to which policy gradient converges to is *deterministic* in line with Theorem 4.5. The panels in the middle and right columns depict the L1-accuracy in the policy space at each iteration which is defined as the average distance between the current policy and the final policy of all 8 agents, i.e., L1-accuracy $= \frac{1}{N} \sum_{i \in \mathcal{N}} |\pi_i - \pi_i^{\text{final}}| = \frac{1}{N} \sum_{i \in \mathcal{N}} \sum_s \sum_a |\pi_i(a \mid s) - \pi_i^{\text{final}}(a \mid s)|$. The results are qualitatively equivalent in both cases (common and non-common learning rates). However, due to the larger step-sizes used by some agents, the algorithm becomes more responsive and exhibits faster convergence in the non-common case.

6   Further Discussion and Conclusion

In this paper, we have presented a number of positive results (both structural and algorithmic) about the performance of independent policy gradient ascent in Markov potential games. Specifically, deterministic Nash policies always exist and independent policy gradient is guaranteed to quickly converge to a Nash policy profile even in the case of finite samples. Given these positive results, a number of interesting open questions emerge.

Open questions.   Price of (Markov) Anarchy. Price of Anarchy (PoA) [16] is a classic notion in normal form games that captures the inefficiency due to the lack of a central authority that can coordinate all agents to implement the social optimum outcome. Formally, it is defined as the ratio of the social cost of the worst Nash equilibrium divided by the cost of the social optimum. PoA has been studied extensively in many classes of games including several classes of potential games for which tight PoA bounds exist (e.g. congestion games [28]). It would be interesting to explore to what extent this type of analysis can be generalized to Markov Potential Games as well as more general classes of Markov Games.

   Stability of Deterministic Policies. When it comes to online learning in normal form potential games, it sometimes possible to prove that the dynamics do not converge to an arbitrary Nash equilibrium but that in fact most initial conditions converge to a deterministic (sometimes referred to also as pure) Nash equilibrium [15, 26, 8, 19]. To produce such equilibrium selection results, standard Lyapunov arguments do not suffice and one needs to apply more advanced techniques such as the Center-Stable-Manifold theorem [17] which would be a fascinating direction for future work in Markov potential games.

   Other Algorithmic Approaches: Softmax Parametrization & Natural Policy Gradient. In [1], the authors show asymptotic convergence to the global optimum to single-agent MDP in the tabular setting with exact gradients for the softmax policy parameterization. Moreover, polynomial convergence rate is shown when additional KL-based entropy regularizer is used, as well as dimension-free convergence to optimum when Natural Policy Gradient is applied. Extending such algorithmic techniques to the case of multi-agent MPGs is a natural direction for future work.

   Global Convergence in other Multi-Agent Markov Games. Recently, there has been intense interest in understanding convergence to Nash policies for different classes of learning dynamics in Markov zero-sum games [11, 38, 40]. Our approach moves in orthogonal direction focusing on MPGs and establishing strong convergence results in these games. A

natural open question is whether and under what conditions can we prove strong convergence guarantees in more general classes of Markov games, possibly by combining tools and techniques from both lines of work.

Regularities beyond Equilibration in Multi-Agent Markov Games. Given the complexities of such multi-agent settings, it is highly unlikely to expect practical algorithms which can always guarantee convergence to equilibrium. This is already the case even for the more restricted setting of normal-form games [36, 2]. Nevertheless, strong guarantees can be shown via, e.g., existence of cyclic/recurrent orbits, invariant functions [20] or strong social welfare guarantees [33]. Whether such results can be extended to Multi-Agent Markov Games is a stimulating direction for future work.

References

[1] A. Agarwal, S. M. Kakade, J. D. Lee, and G. Mahajan. Optimality and Approximation with Policy Gradient Methods in Markov Decision Processes. In J. Abernethy and S. Agarwal, editors, *Proceedings of 33rd Conference on Learning Theory*, volume 125 of *PMLR*, pages 64–66, 2020.

[2] Gabriel P Andrade, Rafael Frongillo, and Georgios Piliouras. Learning in matrix games can be arbitrarily complex. *arXiv preprint arXiv:2103.03405*, 2021.

[3] Dimitri P. Bertsekas. *Dynamic Programming and Optimal Control*. Athena Scientific, 2nd edition, 2000.

[4] I. Bistritz and N. Bambos. Cooperative multi-player bandit optimization. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 2016–2027. Curran Associates, Inc., 2020.

[5] Noam Brown and Tuomas Sandholm. Superhuman ai for multiplayer poker. *Science*, 365(6456):885–890, 2019.

[6] Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Found. Trends Mach. Learn.*, 8(3-4):231–357, 2015.

[7] Lucian Busoniu, Robert Babuska, and Bart De Schutter. A comprehensive survey of multiagent reinforcement learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 38(2):156–172, 2008.

[8] Johanne Cohen, Amélie Héliou, and Panayotis Mertikopoulos. Learning with bandit feedback in potential games. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6372–6381, 2017.

[9] A. Dafoe, Y. Bachrach, G. Hadfield, E. Horvitz, K. Larson, and T. Graepel. Cooperative ai: machines must learn to find common ground. *Nature*, 7857:33–36, 2021.

[10] A. Dafoe, E. Hughes, Y. Bachrach, T. Collins, K. R. McKee, J. Z. Leibo, K. Larson, and T. Graepel. Open Problems in Cooperative AI. *arXiv e-prints*, December 2020.

[11] C. Daskalakis, D.J. Foster, and N. Golowich. Independent Policy Gradient Methods for Competitive Reinforcement Learning. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 5527–5540. Curran Associates, Inc., 2020.

[12] Damek Davis and Dmitriy Drusvyatskiy. Stochastic subgradient method converges at the rate $o(k^{-1/4})$ on weakly convex functions. *CoRR*, abs/1802.02988, 2018.

[13] R. Hassin and M. Haviv. *To queue or not to queue: Equilibrium behavior in queueing systems*. Kluwer Academic Publishers, Boston, USA, 2003.

[14] S. Kakade and J. Langford. Approximately Optimal Approximate Reinforcement Learning. In *Proceedings of the Nineteenth International Conference on Machine Learning*, ICML '02, page 267–274, San Francisco, CA, USA, 2002. Morgan Kaufmann Publishers Inc.

[15] Robert Kleinberg, Georgios Piliouras, and Éva Tardos. Multiplicative updates outperform generic no-regret learning in congestion games. In *ACM Symposium on Theory of Computing (STOC)*, 2009.

[16] E. Koutsoupias and C. Papadimitriou. Worst-case equilibria. In *(STACS)*, pages 404–413. Springer-Verlag, 1999.

[17] Jason D Lee, Ioannis Panageas, Georgios Piliouras, Max Simchowitz, Michael I Jordan, and Benjamin Recht. First-order methods almost always avoid strict saddle points. *Mathematical programming*, 176(1):311–337, 2019.

[18] J. R. Marden. State based potential games. *Automatica*, 48(12):3075–3088, 2012.

[19] Ruta Mehta, Ioannis Panageas, and Georgios Piliouras. Natural selection as an inhibitor of genetic diversity: Multiplicative weights updates algorithm and a conjecture of haploid genetics. In *Innovations in Theoretical Computer Science*, 2015.

[20] Panayotis Mertikopoulos, Christos Papadimitriou, and Georgios Piliouras. Cycles in adversarial regularized learning. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2703–2717. SIAM, 2018.

[21] D. Mguni. Stochastic Potential Games. *arXiv e-prints*, page arXiv:2005.13527, May 2020.

[22] D. Mguni, Y. Wu, Y. Du, Y. Yang, Z. Wang, M. Li, Y. Wen, J. Jennings, and J. Wang. Learning in Nonzero-Sum Stochastic Games with Potentials. *arXiv e-prints*, page arXiv:2103.09284, March 2021.

[23] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, Feb 2015.

[24] D. Monderer and L. S. Shapley. Potential Games. *Games and Economic Behavior*, 14(1):124–143, 1996.

[25] OpenAI. Openai five. openai.com, 2018.

[26] Ioannis Panageas, Georgios Piliouras, and Xiao Wang. Multiplicative weights update as a distributed constrained optimization algorithm: Convergence to second-order stationary points almost always. In *ICML*, 2018.

[27] T. Roughgarden. Intrinsic robustness of the price of anarchy. *J. ACM*, 62(5), November 2015.

[28] Tim Roughgarden and Éva Tardos. How bad is selfish routing? *Journal of the ACM (JACM)*, 49(2):236–259, 2002.

[29] L. S. Shapley. Stochastic games. *PNAS*, 1953.

[30] David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489, Jan 2016.

[31] David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, Timothy Lillicrap, Karen Simonyan, and Demis Hassabis. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, 362(6419):1140–1144, 2018.

[32] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. A Bradford Book, Cambridge, MA, USA, 2018.

[33] Vasilis Syrgkanis, Alekh Agarwal, Haipeng Luo, and Robert E. Schapire. Fast convergence of regularized learning in games. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*, NIPS'15, pages 2989–2997, Cambridge, MA, USA, 2015. MIT Press.

[34] S. Valcarcel Macua, J. Zazo, and S. Zazo. Learning Parametric Closed-Loop Policies for Markov Potential Games. In *International Conference on Learning Representations*, 2018.

[35] Oriol Vinyals, Igor Babuschkin, Wojciech M. Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H. Choi, Richard Powell, Timo Ewalds, Petko Georgiev, Junhyuk Oh, Dan Horgan, Manuel Kroiss, Ivo Danihelka, Aja Huang, Laurent Sifre, Trevor Cai, John P. Agapiou, Max Jaderberg, Alexander S. Vezhnevets, Rémi Leblond, Tobias Pohlen, Valentin Dalibard, David Budden, Yury Sulsky, James Molloy, Tom L. Paine, Caglar Gulcehre, Ziyu Wang, Tobias Pfaff, Yuhuai Wu, Roman Ring, Dani Yogatama, Dario Wünsch, Katrina McKinney, Oliver Smith, Tom Schaul, Timothy Lillicrap, Koray Kavukcuoglu, Demis Hassabis, Chris Apps, and David Silver. Grandmaster level in starcraft ii using multi-agent reinforcement learning. *Nature*, 575(7782):350–354, Nov 2019.

[36] Emmanouil-Vasileios Vlatakis-Gkaragkounis, Lampros Flokas, Panayotis Mertikopoulos, and Georgios Piliouras. No-regret learning and mixed nash equilibria: They do not mix. In *Annual Conference on Neural Information Processing Systems*, 2020.

[37] X. Wang and T. Sandholm. Reinforcement Learning to Play an Optimal Nash Equilibrium in Team Markov Games. In *Proceedings of the 15th International Conference on Neural Information Processing Systems*, NIPS'02, page 1603–1610, Cambridge, MA, USA, 2002. MIT Press.

[38] Chen-Yu Wei, Chung-Wei Lee, Mengxiao Zhang, and Haipeng Luo. Last-iterate convergence of decentralized optimistic gradient descent/ascent in infinite-horizon competitive markov games. *CoRR*, abs/2102.04540, 2021.

[39] K. Zhang, Z. Yang, and T. Başar. Multi-Agent Reinforcement Learning: A Selective Overview of Theories and Algorithms. *arXiv e-prints*, page arXiv:1911.10635, 2019.

[40] Yulai Zhao, Yuandong Tian, Jason D. Lee, and Simon S. Du. Provably efficient policy gradient methods for two-player zero-sum markov games. *CoRR*, abs/2102.08903, 2021.

A Additional Notation and Definitions: Section 2

We first provide some additional notation and definitions that will be used in the proofs.

Q-value and Advantage Functions. Recall from the main part that the value function, $V_s^i : \Pi \to \mathbb{R}$, gives the expected reward of agent $i \in \mathcal{N}$ when $s_0 = s$ and the agents draw their actions, $\mathbf{a}_t = (a_{i,t}, \mathbf{a}_{-i,t})$, at time $t \geq 0$ from policies $\pi = (\pi_i, \pi_{-i})$ and is defined as

$$V_s^i(\pi) := \mathbb{E}_{\tau \sim \pi}\left[\sum_{t=0}^{\infty} \gamma^t r_{i,t} \mid s_0 = s \right].$$

Similarly, we will write $V_\rho^i(\pi) := \mathbb{E}_{s_o \sim \rho}[V_s^i(\pi)]$ to denote the expected value of agent $i \in \mathcal{N}$ under the initial state distribution ρ.

For any state $s \in \mathcal{S}$, the Q-value function $Q_s^i : \mathcal{P} \times \mathcal{A} \to \mathbb{R}$ and the advantage function $A_s^i : \mathcal{P} \times \mathcal{A} \to \mathbb{R}$ of agent $i \in \mathcal{N}$ are defined as

$$Q_s^i(\pi, \mathbf{a}) := \mathbb{E}_{\tau \sim \pi}\left[\sum_{t=0}^{\infty} \gamma^t r_{i,t} \mid s_0 = s, \mathbf{a}_0 = \mathbf{a} \right], \text{ and} \tag{15}$$

$$A_s^i(\pi, \mathbf{a}) := Q_s^i(\pi, \mathbf{a}) - V_s^i(\pi). \tag{16}$$

Discounted State Distribution. It will be useful to define the discounted state visitation distribution $d_{s_0}^\pi(s)$ for $s \in \mathcal{S}$ that is induced by a (joint) policy π as

$$d_{s_0}^\pi(s) := (1 - \gamma) \sum_{t=0}^{\infty} \gamma^t \mathrm{Pr}^\pi(s_t = s \mid s_0), \quad \text{for all } s \in \mathcal{S}. \tag{17}$$

As for the value function, we will also write $d_\rho^\pi(s) = \mathbb{E}_{s_0 \sim \rho}[d_{s_0}^\pi(s)]$ to denote the discounted state visitation distribution when the initial state distribution is ρ.

B Omitted Materials: Section 3

Proposition B.1 (Separability of Value Functions and Equality of Derivatives). *Let $\mathcal{G} = (\mathcal{S}, \mathcal{N}, \mathcal{A} = \{\mathcal{A}_i\}_{i \in \mathcal{N}}, P, R, \rho)$ be a Markov Potential Game (MPG) with potential Φ_s, for $s \in S$. Then, for the value function $V_s^i, s \in \mathcal{S}$ of each agent $i \in \mathcal{N}$, the following hold*

P1. *Separability of Value Functions: there exists a function $U_s^i : \Delta(\mathcal{A}_{-i})^S \to \mathbb{R}$ such that for each joint policy profile $\pi = (\pi_i, \pi_{-i}) \in \Delta(\mathcal{A})^S$, we have $V_s^i(\pi) = \Phi_s(\pi) + U_s^i(\pi_{-i})$.*

P2. *Equality of Derivatives: the partial derivatives of agent i's value function V_s^i coincide with the partial derivatives of the potential Φ_s that correspond to agent i's parameters, i.e.,*

$$\partial_{x_{i,s,a}} V_s^i(\pi) = \partial_{x_{i,s,a}} \Phi_s(\pi), \quad \text{for all } i \in \mathcal{N} \text{ and all } s \in \mathcal{S}.$$

Proof. To obtain P1, consider any 3 arbitrary policies for agent i, notated by $\pi_i, \pi_i', \pi_i'' \in \Delta(\mathcal{A}_i)^S$. Then, by the definition of MPGs, we have that

$$\Phi_s(\pi_i, \pi_{-i}) - \Phi_s(\pi_i', \pi_{-i}) = V_s^i(\pi_i, \pi_{-i}) - V_s^i(\pi_i', \pi_{-i}),$$
$$\Phi_s(\pi_i, \pi_{-i}) - \Phi_s(\pi_i'', \pi_{-i}) = V_s^i(\pi_i, \pi_{-i}) - V_s^i(\pi_i'', \pi_{-i}).$$

for every starting state $s \in \mathcal{S}$. This implies that we can write $V_s^i(\pi_i, \pi_{-i})$ as both

$$V_s^i(\pi_i, \pi_{-i}) = \Phi_s(\pi_i, \pi_{-i}) - \Phi_s(\pi_i', \pi_{-i}) + V_s^i(\pi_i', \pi_{-i})$$
$$V_s^i(\pi_i, \pi_{-i}) = \Phi_s(\pi_i, \pi_{-i}) - \Phi_s(\pi_i'', \pi_{-i}) + V_s^i(\pi_i'', \pi_{-i})$$

Thus, we have that $-\Phi_s(\pi_i', \pi_{-i}) + V_s^i(\pi_i', \pi_{-i}) = -\Phi_s(\pi_i'', \pi_{-i}) + V_s^i(\pi_i'', \pi_{-i})$ for any arbitrary pair of policies π_i' and π_i'' for agent i, implying that agent i's policy has no impact on these terms. Accordingly, we can express them as

$$U_s^i(\pi_{-i}) := -\Phi_s(\pi_i', \pi_{-i}) + V_s^i(\pi_i', \pi_{-i}) = -\Phi_s(\pi_i'', \pi_{-i}) + V_s^i(\pi_i'', \pi_{-i}),$$

where $U_s^i(\pi_{-i})$ is a function that does not depend on the policy of agent i. Thus, we can express the utility function of any agent i in a MPG as

$$V_s^i(\pi) = \Phi_s(\pi) + U_s^i(\pi_{-i}).$$

as claimed. To obtain P2, we use P1 for a vector x_i parameterizing π_i, and obtain that

$$\partial_{x_{i,a}} V^i(\pi) = \partial_{x_{i,a}} \Phi(\pi) + 0$$

for any coordinate $x_{i,a}$ with $a \in A_i$ of x_i, from which we can see that our claim is true. $\square$

Note that P1 serves as a characterization of MPGs. Namely, a multi-agent MDP is a MPG if and only if the value function of each agent $i \in \mathcal{N}$ can be decomposed in a term that is common for all players (potential function) and in a term that may be different for each agent $i \in \mathcal{N}$ but which depends only on the actions of all agents other than i. This property carries over from normal form (single state) potential games. Also note that both properties, P1 and P2, hold for any (differentiable for P2) policy parameterization and not only for the direct one that we use here.

Proof of Theorem 3.1. Let Φ be the potential function of $\mathcal{G}$. Since the space $\Delta(\mathcal{A})^S = \Delta(\mathcal{A}_1)^S \times ... \times \Delta(\mathcal{A}_n)^S$ is compact and Φ is continuous, Φ has a global maximum $\Phi_{\max}$. Let $(\pi_1^*, ..., \pi_n^*)$ denote a global maximizer, i.e., a joint policy profile at which $\Phi_{\max}$ is attained. By the Definitions of MPGs and Nash policies, this implies, in particular, that $(\pi_1^*, ..., \pi_n^*)$ is a Nash policy, since

$$0 < \Phi_s(\pi_i^*, \pi_{-i}^*) - \Phi_s(\pi_i, \pi_{-i}^*) = V_s^i(\pi_i^*, \pi_{-i}^*) - V_s^i(\pi_i, \pi_{-i}^*), \qquad (*)$$

for all $i \in \mathcal{N}, s \in \mathcal{S}$ and all policies $\pi_i \in \Delta(\mathcal{A}_i)^S$. If $\pi_1^*, ..., \pi_n^*$ are all deterministic we are done. So, we may assume that there exists an $i \in \mathcal{N}$ so that π_i^* is randomized and consider the MDP $\mathcal{G}'$ in which the policy of all agents other than i has been fixed to π_{-i}^*. $\mathcal{G}'$ is a single agent MDP with the same states as $\mathcal{G}$, the same actions and rewards for agent i and transition probabilities that are determined by the joint distribution of the environment and the joint policy of all agents other than i. As a single agent MDP, this setting has a deterministic optimal policy, say $\tilde{\pi}_i$, for agent i. Thus, it holds that

$$V_s^i(\tilde{\pi}_i, \pi_{-i}^*) \leq V_s^i(\pi_i^*, \pi_{-i}^*) \leq V_s^i(\tilde{\pi}_i, \pi_{-i}^*),$$

where the first inequality follows from the fact that π^* is a Nash policy and the second from the optimality of $\tilde{\pi}_i$. It follows that

$$V_s^i(\tilde{\pi}_i, \pi_{-i}^*) = V_s^i(\pi_i^*, \pi_{-i}^*),$$

i.e., the payoff of agent i at $(\tilde{\pi}_i, \pi_{-i}^*)$ is the same as in (π_i^*, π_{-i}^*). Hence, by the definition of the potential function, we have that

$$0 = V_s^i(\tilde{\pi}_i, \pi_{-i}^*) - V_s^i(\pi_i^*, \pi_{-i}^*) = \Phi_s(\tilde{\pi}_i, \pi_{-i}^*) - \Phi_s(\pi_i^*, \pi_{-i}^*),$$

which implies that

$$\Phi_s(\tilde{\pi}_i, \pi_{-i}^*) = \Phi_s(\pi_i^*, \pi_{-i}^*) = \Phi_{\max}.$$

Thus, $(\tilde{\pi}_i, \pi_{-i}^*)$ is also a global maximizer of Φ which implies that $(\tilde{\pi}_i, \pi_{-i}^*)$ is a Nash policy by the same reasoning as in equation $(*)$. Note that the value of all players other than i may not be the same at the joint policy profile $(\tilde{\pi}_i, \pi_{-i}^*)$ as it is in (π_i^*, π_{-i}^*). However, what we need for our purpose is that this step reduces the number of randomized policies by one and that it retains the value of the potential function invariant at its global maximum (which ensures that the ensuing policy profile is also a Nash policy). By iterating this process until π_j^* becomes deterministic for all agents $j \in \mathcal{N}$, we obtain the claim. $\qquad \square$

Proof of Proposition 3.2. The proof is constructive and proceeds by finding the potential function, $\Phi_s, s \in \mathcal{S}$ in both cases, C1-C2. Since the individual rewards of the agents at each state $s \in \mathcal{S}$ are captured by a potential function ϕ_s, then for the reward, $R_i(s, \mathbf{a})$ of each agent i at the action profile $\mathbf{a}$, it holds that

$$R_i(s, \mathbf{a}) = \phi_s(\mathbf{a}) + u_s^i(\mathbf{a}_{-i}), \qquad (18)$$

where $u_s^i : \Delta(\mathcal{A}_{-i}) \to \mathbb{R}$ is a function that does not depend on the actions of agent i in any way. Thus, we may write the value function of each agent $i \in \mathcal{N}$ as

$$V_s^i(\pi) = \mathbb{E}_{\tau \sim \pi} \left[\sum_{t=0}^{\infty} \gamma^t R_i(s_t, \mathbf{a}_t) \mid s_0 = s \right]$$

$$= \mathbb{E}_{\tau \sim \pi} \left[\sum_{t=0}^{\infty} \gamma^t \left(\phi_s(\mathbf{a}_t) + u_s^i(\mathbf{a}_{-i,t}) \right) \mid s_0 = s \right]$$

$$= \mathbb{E}_{\tau \sim \pi} \left[\sum_{t=0}^{\infty} \gamma^t \phi_s(\mathbf{a}_t) \mid s_0 = s \right] + \mathbb{E}_{\tau \sim \pi} \left[\sum_{t=0}^{\infty} \gamma^t u_s^i(\mathbf{a}_{-i,t}) \mid s_0 = s \right] \qquad (\star)$$

where $\tau \sim \pi$ is the random trajectory generated by policy π. To show that $\mathcal{G}$ is a MPG, it suffices to show that the value function of each agent $i \in \mathcal{N}$ can be decomposed in a term that is common for all agents (and which may depend on the actions of agent $i \in \mathcal{N}$) and in a term that does not depend (in any way) in the actions of agent i (dummy term), cf Proposition B.1. The first term in expression $(\star)$, i.e., $\mathbb{E}_{\tau \sim \pi} [\sum_{t=0}^{\infty} \gamma^t \phi_s(\mathbf{a}_t) \mid s_0 = s]$, depends on the actions of all players and is common for all agents $i \in \mathcal{N}$ (and is thus, a good candidate for the potential function). The second term in expression $(\star)$, i.e., $\mathbb{E}_{\tau \sim \pi} [\sum_{t=0}^{\infty} \gamma^t u_s^i(\mathbf{a}_{-i,t}) \mid s_0 = s]$, does not depend on player i via the payoffs $u_s^i(\mathbf{a}_{-i,t})$, but, in general, it does depend on player i via the transitions $\tau \sim \pi$. The two cases in the statement of Proposition 3.2 ensure precisely that this term is either independent of agent i, in which case it is a dummy term for agent i or that is also common for all players, in which case it can be included in the potential function. Specifically, we have that

C1. If the transitions do not depend on the action of the players, we have that $\tau \sim P$, where P is an exogenously given distribution function (state-wise). In this case, we have that

$$\Phi_s(\pi) := \mathbb{E}_{\tau \sim \pi} \left[\sum_{t=0}^{\infty} \gamma^t \phi_s(\mathbf{a}_t) \mid s_0 = s \right]$$

is a potential function and $U_s^i(\pi_{-i}) := \mathbb{E}_{\tau \sim \pi} [\sum_{t=0}^{\infty} \gamma^t u_s^i(\mathbf{a}_{-i}) \mid s_0 = s]$ is a dummy term that does not depend (in any way) on the policy of agent $i \in \mathcal{N}$. For a rigorous argument, let $\alpha_{\mathbf{a},s}^t(\pi_i, \pi_i') = \Pr[\mathbf{a}_t = \mathbf{a} \mid s_t = s] - \Pr[\mathbf{a}_t' = \mathbf{a} \mid s_t' = s]$ for any policies $\pi = (\pi_i, \pi_{-i})$ and $\pi' = (\pi_i', \pi_{-i})$. By (crucially) using the fact that $\Pr^{\pi}[s_t = s] = \Pr^{\pi'}[s_t' = s]$ according to our assumption that P does not depend on actions, we then have that

$$\mathbb{E}_{\tau \sim \pi} \left[\sum_{t=0}^{\infty} \gamma^t u_s^i(\mathbf{a}_{-i,t}) \mid s_0 = s \right] - \mathbb{E}_{\tau \sim \pi'} \left[\sum_{t=0}^{\infty} \gamma^t u_s^i(\mathbf{a}_{-i,t}') \mid s_0 = s \right] =$$

$$= \sum_s \Pr[s_t = s] \left(\sum_{\mathbf{a}} \alpha_{\mathbf{a},s}^t(\pi, \pi_i') u_s^i(\mathbf{a}) \right)$$

Now note that we must have $\sum_{\mathbf{a}} \alpha_{\mathbf{a},s}^t(\pi_i, \pi_i') = 0$. Intuitively, for whatever probability went up for an action when agent i went from π_i to π_i', an equal mass of probability must have gone down for other actions. Let $A^+ = \{\mathbf{a} \mid \alpha_{\mathbf{a},s}^t(\pi_i, \pi_i') > 0\}$ and $A^- = \{\mathbf{a} \mid \alpha_{\mathbf{a},s}^t(\pi_i, \pi_i') < 0\}$.

26

Then this implies we can find some number M of pairs $(\mathbf{a_m^+}, \mathbf{a_m^-}) \in A^+ \times A^-$, with coefficients β_m, such that $\sum_m \beta_m = \sum_{\mathbf{a} \in A^+} \alpha_{\mathbf{a},s}^t(\pi_i, \pi_i') = -\sum_{\mathbf{a} \in A^-} \alpha_{\mathbf{a},s}^t(\pi_i, \pi_i')$, and such that we have

$$\sum_s \Pr[s_t = s] \left(\sum_{\mathbf{a}} \alpha_{\mathbf{a},s}^t(\pi, \pi_i') u_s^i(\mathbf{a}) \right) = \sum_s \Pr[s_t = s] \sum_m \beta_m \left((u_s^i(\mathbf{a_m^+}) - u_s^i(\mathbf{a_m^-})) \right)$$
$$= \sum_s \Pr[s_t = s] \sum_m \beta_m \left((\phi_s(\mathbf{a_m^+}) - \phi_s(\mathbf{a_m^-})) \right)$$

where the last equality uses the fact that the game at state s is a potential game. Now, the right hand side of the last expression is equal to the difference of the common terms, i.e.,

$$\sum_s \mathbb{P}[s_t = s] \sum_m \beta_m \left((\phi_s(\mathbf{a_m^+}) - \phi_s(\mathbf{a_m^-})) \right) =$$
$$= \mathbb{E}_{\tau \sim \pi} \left[\sum_{t=0}^\infty \gamma^t \phi_s(\mathbf{a}_{-i,t}) \mid s_0 = s \right] - \mathbb{E}_{\tau \sim \pi'} \left[\sum_{t=0}^\infty \gamma^t \phi_s(\mathbf{a}_{-i,t}') \mid s_0 = s \right]$$

which, thus implies that

$$\Phi_s(\pi_i, \pi_{-i}) - \Phi_s(\pi_i', \pi_{-i}) = V_s^i(\pi_i, \pi_{-i}) - V_s^i(\pi_i', \pi_{-i})$$

as claimed.

C2. If the dummy terms for each player $i \in \mathcal{N}$ are equal across states, i.e., if there exists a function $u^i : \Delta(\mathcal{A}_{-i})^S \to \mathbb{R}$ so that $u_s^i(\mathbf{a}_{-i}) = u^i(\mathbf{a}_{-i})$ for all $s \in \mathcal{S}$, then it holds that the term $\mathbb{E}_{\tau \sim \pi} \left[\sum_{t=0}^\infty \gamma^t u_s^i(\mathbf{a}_{-i}) \mid s_0 = s \right]$ is equal for all trajectories $\tau \sim \pi$. Hence, this term does not in the actions of player i and the same decomposition as in condition C1 applies.

Summing up, in both cases, C1-C2, $\mathcal{G}$ is an MPG as claimed.

$\square$

Note that the proof of the (trivial) case in which the instantaneous rewards of all agents $i \in \mathcal{N}$ are equal at each state $s \in \mathcal{S}$ (cf. Remark 2) is similar. In this case, it is immediate to see that the instantaneous rewards are precisely given by the potential function at that state, i.e., it holds that $R_i(s, \mathbf{a}) = \phi_s(\mathbf{a})$ for all $i \in \mathcal{N}$ and all $s \in \mathcal{S}$. In this case, it holds that $u_s^i(\mathbf{a}_{-i}) \equiv 0$ for all $i \in \mathcal{N}$ and all $s \in \mathcal{S}$ and hence,

$$\Phi_s(\pi) := \mathbb{E}_{\tau \sim \pi} \left[\sum_{t=0}^\infty \gamma^t \phi_s(\mathbf{a}) \mid s_0 = s \right]$$

is a potential function for G, and the dummy terms are all equal to 0, i.e., $U_s^i(\pi_{-i}) \equiv 0$.

B.1 Examples

Example 2 (Continued). At each state, $s \in \{0, 1\}$, the agents' payoffs, (R_s^1, R_s^2), form a potential game (at that state), and are given as follows

$$\text{State } 0: \quad (R_0^1, R_0^2) = \begin{array}{c} \\ 0 \\ 1 \end{array}\begin{array}{c} 0 \qquad\quad 1 \\ \left(\begin{array}{cc} 5, \; 2 & -1, -2 \\ -5, -4 & 1, \; 4 \end{array} \right) \end{array}, \quad \text{with potential } \Phi_0 = \begin{pmatrix} 4 & 0 \\ -6 & 2 \end{pmatrix},$$

$$\text{State } 1: \quad (R_1^1, R_1^2) = (0, 0), \quad \text{with potential } \Phi_1 = 0.$$

In this MDP, agents need only to select an action at state s_0. Thus, we will denote a policy, π_1, of agent 1 by $\pi_1 = (p, 1 - p)$ where $p \in [0, 1]$ is the probability with which agent A selects action 0 at state s_0. Similarly, we will denote a policy, π_2, of agent B by $\pi_2 = (q, 1 - q)$ where $q \in [0, 1]$ is the probability with which agent B selects action 0 at state s_0. Moreover, we will slightly abuse notation and write

$$R_0^i(\pi) = R_0^i(\pi_1, \pi_2) = \pi_1^\top R_0^i \pi_2 = [p, 1 - p] R_0^i [q, 1 - q]^\top.$$

We also assume that the horizon is infinite and there is a discount factor $\gamma \in [0, 1)$. Accordingly, we can calculate the value functions $V_0^i(\pi_1, \pi_2)$ of agents $i = A, B$ starting from state s_0 as follows,

$$V_0^i(\pi) = R_0^i(\pi) + \gamma pq V_0^i(\pi) - \gamma(1 - pq) \times 0$$

which yields the solution

$$V_0^i(\pi) = \frac{R_0^i(\pi)}{1 - \gamma pq}, \quad \text{for } i = A, B.$$

Next, we use the Performance Difference Lemma (Lemma 3.2 by [1]) to determine the difference in the value between two different policies. We will do this for agent 1 (the calculation is similar for agent 2: we use here 1 for agent 1 and 2 for agent B). For a policy $\pi = (\pi_1, \pi_2)$, we have at state s_0 that

$$\mathbb{E}_{a \sim \pi_1(\cdot|s_0)}[A_0^1(\pi', \mathbf{a})] = p A_0^1(\pi', 0, a_2) + (1 - p) A^1(\pi', 1, a_2)$$

$$= p \left[R_0^1(0, \pi_2) + \gamma q V_0(\pi') - V_0(\pi') \right] + (1 - p) \left[R_0^1(1, \pi_2) + 0 - V_0(\pi') \right]$$

$$= p R_0^1(0, \pi_2) + (1 - p) R_0^1(0, \pi_2) - (1 - \gamma pq) V_0(\pi')$$

$$= R_0^1(\pi) - (1 - \gamma pq) V_0(\pi').$$

At state s_1, there is only one available action for each agent which yields a payoff of 0. Thus,

$$\mathbb{E}_{a \sim \pi_1(\cdot|s_1)}[A_1^1(\pi', \mathbf{a})] = 0.$$

Moreover, concerning the discounted visitation distribution, we have that

$$d_0^\pi(s_0) = (1 - \gamma) \sum_{t=0}^{\infty} \gamma^t \mathrm{Pr}^\pi(s_t = s_0 \mid s_0) = (1 - \gamma) \left[1 + \gamma pq + (\gamma pq)^2 + \ldots \right] = \frac{1 - \gamma}{1 - \gamma pq},$$

and $d_0^\pi(s_1) = 1 - d_0^\pi(s_0) = \frac{\gamma(1-pq)}{1-\gamma pq}$. Thus, using all the above, we have that

$$V_0(\pi) - V_0(\pi') = \frac{1}{1-\gamma}\left[\frac{1-\gamma}{1-\gamma pq}\cdot(R_0^1(\pi) - (1-\gamma pq)V_0(\pi')) + \frac{\gamma(1-pq)}{1-\gamma pq}\cdot 0\right]$$

$$= \frac{R_0^1(\pi)}{1-\gamma pq} - V_0(\pi') = V_0(\pi) - V_0(\pi').$$

which shows that our initial calculations conform with the outcome specified by the Performance Difference Lemma.

Finally, a direct calculation shows that $\Phi_s = \phi_s$ for $s = 0, 1$ is a valid potential function for which the MDP is an *ordinal* MPG.

Example 3 (Continued). At state s_0, we consider the game with action sets $A_1(s_0) = A_2(s_0) = \{H, T\}$ and (instantaneous) payoffs

$$R_1(s_0, a_1, a_2) = \begin{array}{c} \\ H \\ T \end{array}\begin{array}{cc} H & T \\ \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} \end{array} \quad \text{and} \quad R_2(s_0, a_1, a_2) = \begin{array}{c} \\ H \\ T \end{array}\begin{array}{cc} H & T \\ \begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix} \end{array},$$

where a_1 denotes the action of agent 1 and a_2 the action of agent 2 (agent 1 selects rows and agent 2 selects columns in both matrices). This is a constant sum game (equivalent to zero-sum) and hence, it is not an (ordinal) potential game. Apart from the instantaneous rewards, agents' actions at s_0 induce a deterministic transition to a state in which the only available actions to the agents are precisely the actions that they chose at state s_0 and their instantaneous rewards at this state are the rewards of the other agent at s_0. In particular, there are four possible transitions to states s_{ab} with $a, b \in \{H, T\}$, with action sets and instantaneous rewards given by

$$A_1(s_{ab}) = \{a\}, A_2(s_{ab}) = \{b\}, \quad R_1(s_{ab}, a, b) = R_2(s_0, b, a), \ R_2(s_{ab}, a, b) = R_1(s_0, b, a),$$

for agents 1 and 2, respectively. Note that the visitation probability of this states is equal to the visitation probability of state s_0. After visiting one of these states, the MDP transitions to state s_1 which is a potential game, with potential function given by

$$\Phi_1 = \begin{array}{c} \\ L \\ R \end{array}\begin{array}{cc} L & R \\ \begin{pmatrix} 4 & 3 \\ 3 & 0 \end{pmatrix} \end{array}$$

As mentioned above, the game in state s_0 does not admit a potential function. However, the joined rewards RJ_1, RJ_2 of agents 1 and 2 which result from selecting an action profile $(a, b) \in H, T^2$ at s_0 and then traversing both s_0 and the ensuing s_{ab} (part included in the dotted rectangle in Figure 3), do admit a potential function. The potential function in this case is the sum of agents' rewards and is given by

$$\Phi_{0ab} = \begin{array}{c} \\ H \\ T \end{array}\begin{array}{cc} H & T \\ \begin{pmatrix} 1-1 & 1-1 \\ 1-1 & 1-1 \end{pmatrix} \end{array} = \begin{array}{c} \\ H \\ T \end{array}\begin{array}{cc} H & T \\ \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} \end{array}.$$

29

Let $\pi_1 = (\mathbf{x}_0, \mathbf{x}_1)$ denote a policy of agent 1. Here $\mathbf{x}_0 = (x_0, 1 - x_0)$, where $x_0 \in [0, 1]$ is the probability with which agent 1 chooses action H at state s_0. Similarly, $\mathbf{x}_1 = (x_1, 1 - x_1)$ where $x_1 \in [0, 1]$ is the probability with which agent 1 chooses action L at state s_1. At states $s_{ab}, a, b \in \{H, T\}^2$, agents only have one action to choose from, so this choice is eliminated from their policy representation. Similarly, we represent a policy of agent 2 by $\pi_2 = (\mathbf{y}_0, \mathbf{y}_1)$ with $y_0, y_1 \in [0, 1]$. Let also

$$p_0 := p_0(\pi_1, \pi_2) := \Pr(s_{t+1} = s_0 \mid s_t = s_1, \pi_1, \pi_2), \tag{19}$$

In the general case, p_0, i.e., the transition probability from s_1 to s_0, may depend on the actions of the agents or it may be completely exogenous (i.e., constant with respect to agents' actions). If we write

$$p_0(a_1, a_2) := \Pr(s_{t+1} = s_0 \mid s_t = s_1, a_1, a_2), \quad \text{for } a_1, a_2 \in \{L, R\},$$

to denote the probability of transitioning from state s_1 to state s_0 given that the agents chose actions $a_1, a_2 \in \{L, R\}$ at state s_1, then we can write p_0 as

$$p_0 = \mathbb{E}_{(a_1, a_2) \sim (\pi_1, \pi_2)}[p_0(a_1, a_2)] = \sum_{(a_1, a_2) \in \{L, R\}^2} \Pr(a_1, a_2 \mid \pi_1, \pi_2) \cdot p_0(a_1, a_2)$$

$$= x_1 y_1 p_0(L, L) + x_1(1 - y_1) p_0(L, R) + (1 - x_1) y_1 p_0(R, L) + (1 - x_1)(1 - y_1) p_0(R, R). \tag{20}$$

Using this notation, we can now proceed to compute the value function of each state of the MDP in Figure 3. Since the value of states $s_{a,b}, a, b \in H, T^2$ is equal to a constant reward plus the value of state s_1 (discounted by γ), it suffices to calculate the value for states s_0 and s_1. We have that

$$V_0^1(\pi_1, \pi_2) = \mathbf{x}_0 R_0^1 \mathbf{y}_0 + \gamma \left(\mathbf{x}_0 \left(R_0^2 / \gamma \right) \mathbf{y}_0 \right) + \gamma^2 V_1^1(\pi_1, \pi_2)$$
$$V_1^1(\pi_1, \pi_2) = \mathbf{x}_1 R_1^1 \mathbf{y}_1 + \gamma \left[p_0 V_0^1(\pi_1, \pi_2) + (1 - p_0) V_1^1(\pi_1, \pi_2) \right],$$

which after some trivial calculations yield

$$V_0^1(\pi_1, \pi_2) = \mathbf{x}_0 \left(R_0^1 + R_0^1 \right) \mathbf{y}_0 + \gamma^2 V_1^1(\pi_1, \pi_2)$$
$$V_1^1(\pi_1, \pi_2) = \frac{1}{1 - \gamma(1 - p_0)} \left[\mathbf{x}_1 R_1^1 \mathbf{y}_1 + \gamma p_0 V_0^1(\pi_1, \pi_2) \right].$$

This is a system of 2 equations in the 2 unknown quantities, $V_0^1(\pi_1, \pi_2)$ and $V_1^1(\pi_1, \pi_2)$. Solving for these two quantities, yields the unique solution

$$V_0^1(\pi_1, \pi_2) = \frac{1}{1 - \gamma(1 - p_0) - \gamma^3 p_0} \left[(1 - \gamma(1 - p_0)) \mathbf{x}_0 \left(R_0^1 + R_0^2 \right) \mathbf{y}_0 + \gamma^2 \mathbf{x}_1 R_1^1 \mathbf{y}_1 \right].$$
$$V_1^1(\pi_1, \pi_2) = \frac{1}{1 - \gamma(1 - p_0) - \gamma^3 p_0} \left[\gamma p_0 \mathbf{x}_0 \left(R_0^1 + R_0^2 \right) \mathbf{y}_0 + \mathbf{x}_1 R_1^1 \mathbf{y}_1 \right].$$

In the case that p_0 is a constant with respect to π_1, π_2, then both value functions are of the form

$$V_s^i(\pi_1, \pi_2) = c_1(s) \cdot \mathbf{x}_0 \left(R_0^1 + R_0^2 \right) \mathbf{y}_0 + c_2(s) \cdot \mathbf{x}_1 R_1^i \mathbf{y}_1, \quad \text{for } s \in \{s_0, s_1\}, \text{ and } i = \{1, 2\},$$

30

where $c_1(s), c_2(s) > 0$ are appropriate constants that depend only the state $s \in \{s_0, s_1\}$ and on agents $1, 2$. Since the game at s_1 is a potential game, with potential function given by a 2×2 matrix Φ_1, it is immediate to infer that

$$V_s^i(\pi_1', \pi_2) - V_s^i(\pi_1, \pi_2) = \Phi_s(\pi_1', \pi_2) - \Phi_s(\pi_1, \pi_2), \text{ for } s \in \{s_0, s_1\},$$

with

$$\Phi_s(\pi_1, \pi_2) := c_1(s)\, \mathbf{x}_0 \left(R_0^1 + R_0^2 \right) \mathbf{y}_0 + c_2(s) \cdot \mathbf{x}_1 \Phi_1 \mathbf{y}_1, \text{ for } s \in \{s_0, s_1\}.$$

However, if p_0 depends on the actual policies of agents 1 and 2, cf. equation (19), then it is not immediate to determine a potential (or even to decide whether a (exact) potential exists or not).

Remark 5. Several elements of Example 3 have been selected in the sake of simplicity and are not necessary for the main takeaway, i.e., that there are MDP that are not potential at some states but which are MPGs. First, the transitions from s_0 to the states s_{ab} need not be deterministic. To see this, let $q \in (0, 1)$ and assume that if the agents select actions H, T is s_0, then the process transitions with probability q to a state s_{HT} with rewards $(-1, 1)/q\gamma$ and with probability $(1 - q)$ to a state s'_{HT} with rewards $(1, -1)/(1 - q)\gamma$. The rest remains the same. Accordingly, the expected reward for agent 1 after (H, T) has been selected in s_0 is the same as in the current format.

Second, the construction with states s_0 and $s_{ab}, (a, b) \in H, T^2$ is not the only one that leads to such an example. Another very common instance occurs in the case of *aliasing* between s_0 and states s_{ab}, i.e., when the agents cannot tell these states apart. The intuition which carries over from the currently presented example is that the roles of the agents are essentially reversed between the two states but the agents do not know (from the observable features) in which state they are. Thus, any valid policy, selects the same action in both states leading to the same situation as in the presented example.

Finally, if the horizon is finite, then the instantaneous rewards in states s_{ab} still work if we eliminate the scaling factor (here γ). Thus, the construction works in both episodic and continuing settings.

C Omitted Materials: Section 4

Proof of Lemma 4.1. Observe that for any set $\mathcal{X} \subseteq \mathbb{R}^n$, it holds that

$$P_{\mathcal{X}}(y) = \operatorname*{argmin}_{x \in \mathcal{X}} \|x - y\|_2^2.$$

Thus,

$$P_{\Delta(\mathcal{A})^S}(y) = \operatorname*{argmin}_{x \in \Delta(\mathcal{A})^S} \|x - \pi'\|_2^2 = \operatorname*{argmin}_{x_1 \in \Delta(\mathcal{A}_1)^S, \ldots, x_n \in \Delta(\mathcal{A}_n)^S} \sum_{i=1}^n \|x_i - \pi_i'\|_2^2$$

$$= \sum_{i=1}^n \operatorname*{argmin}_{x_i \in \Delta(\mathcal{A}_i)^S} \|x_i - \pi_i'\|_2^2 = (P_{\Delta(\mathcal{A}_1)^S}(\pi_1'), \ldots, P_{\Delta(\mathcal{A}_n)^S}(\pi_n')). \qquad \square$$

To prove Lemma 4.3, we will use a multi-agent version of the Performance Difference Lemma (cf. [1] for a single agent and [11] for two agents).

Lemma C.1 (Multi-agent Performance Difference Lemma). *Consider an n-agent MDP $\mathcal{G}$ and fix an agent $i \in \mathcal{N}$. Then, for any policies $\pi = (\pi_i, \pi_{-i}), \pi' = (\pi'_i, \pi_{-i}) \in \Pi$ and any distribution $\rho \in \Delta(S)$, it holds that*

$$V^i_\rho(\pi) - V^i_\rho(\pi') = \frac{1}{1-\gamma} \mathbb{E}_{s \sim d^\pi_\rho} \mathbb{E}_{a_i \sim \pi_i(\cdot|s)} \mathbb{E}_{\mathbf{a}_{-i} \sim \pi_{-i}(\cdot|s)} \left[A^i_s(\pi', a_i, \mathbf{a}_{-i}) \right],$$

where $\mathbf{a}_{-i} \sim \pi_{-i}(\cdot \mid s)$ denotes the action profile of all agents other than i that is drawn from the product distribution induced by their policies $\pi_{-i} = (\pi_j)_{j \neq i \in \mathcal{N}} \in \Pi_{-i}$.

Proof. For any initial state $s \in \mathcal{S}$ and joint policies $\pi = (\pi_i, \pi_{-i}), \pi' = (\pi'_i, \pi_{-i}) \in \Pi$, it holds that

$$
\begin{aligned}
V^i_s(\pi) - V^i_s(\pi') &= \mathbb{E}_{\tau \sim \pi} \left[\sum_{t=0}^\infty \gamma^t r_{i,t} \mid s_0 = s \right] - V^i_s(\pi') \\
&= \mathbb{E}_{\tau \sim \pi} \left[\sum_{t=0}^\infty \gamma^t \left(r_{i,t} - V_{s_t}(\pi') + V_{s_t}(\pi') \right) \mid s_0 = s \right] - V^i_s(\pi') \\
&= \mathbb{E}_{\tau \sim \pi} \left[\sum_{t=0}^\infty \gamma^t \left(r_{i,t} - V_{s_t}(\pi') + \gamma V_{s_{t+1}}(\pi') \right) \mid s_0 = s \right] \\
&= \mathbb{E}_{\tau \sim \pi} \left[\sum_{t=0}^\infty \gamma^t \left(r_{i,t} + \gamma \mathbb{E} \left[V_{s_{t+1}}(\pi') \mid s_t, a_{i,t}, \mathbf{a}_{-i,t} \right] - V_{s_t}(\pi') \right) \mid s_0 = s \right] \\
&= \mathbb{E}_{\tau \sim \pi} \left[\sum_{t=0}^\infty \gamma^t A^i_{s_t}(\pi', \mathbf{a}_t) \mid s_0 = s \right] \\
&= \frac{1}{1-\gamma} \mathbb{E}_{s' \sim d^\pi_\rho} \mathbb{E}_{a_i \sim \pi_i(\cdot|s')} \mathbb{E}_{\mathbf{a}_{-i} \sim \pi_{-i}(\cdot|s')} \left[A^i_{s'}(\pi', \mathbf{a}) \right].
\end{aligned}
$$

Taking expectation over the states $s \in \mathcal{S}$ with respect to the distribution $\rho \in \Delta(\mathcal{S})$ yields the result. $\qquad\square$

Proof of Lemma 4.3. Fix an agent $i \in \mathcal{N}$ and let $\pi = (\pi_i, \pi_{-i}), \pi^* = (\pi^*_i, \pi_{-i}) \in \Pi = \Delta(\mathcal{A})^S$. By the definition of MPGs (cf. Definition 2), it holds that

$$V^i_\rho(\pi^*) - V^i_\rho(\pi) = \Phi_\rho(\pi^*) - \Phi_\rho(\pi).$$

Thus, using the multi-agent version of the Performance Difference Lemma (cf. Lemma C.1),

we have for any distribution $\mu \in \Delta(\mathcal{S})$ that

$$\Phi_\rho(\pi^*) - \Phi_\rho(\pi) = V_\rho^i(\pi^*) - V_\rho^i(\pi)$$

$$= \frac{1}{1-\gamma} \mathbb{E}_{s \sim d_\rho^{\pi^*}} \mathbb{E}_{a_i \sim \pi_i^*(\cdot|s)} \mathbb{E}_{\mathbf{a}_{-i} \sim \pi_{-i}^*(\cdot|s)} \left[A_s^i(\pi, a_{i,t}, \mathbf{a}_{-i,t}) \right]$$

$$\leq \frac{1}{1-\gamma} \max_{\pi_i' \in \Pi_i} \left\{ \sum_{s \in \mathcal{S}} d_\rho^{\pi^*}(s) \, \mathbb{E}_{a_i \sim \pi_i'(\cdot|s)} \mathbb{E}_{\mathbf{a}_{-i} \sim \pi_{-i}^*(\cdot|s)} \left[A_s^i(\pi, a_i, \mathbf{a}_{-i}) \right] \right\},$$

$$= \frac{1}{1-\gamma} \max_{\pi_i' \in \Pi_i} \left\{ \sum_{s \in \mathcal{S}} \frac{d_\rho^{\pi^*}(s)}{d_\mu^\pi(s)} d_\mu^\pi(s) \, \mathbb{E}_{a_i \sim \pi_i'(\cdot|s)} \mathbb{E}_{\mathbf{a}_{-i} \sim \pi_{-i}^*(\cdot|s)} \left[A_s^i(\pi, a_i, \mathbf{a}_{-i}) \right] \right\},$$

$$\leq \frac{1}{1-\gamma} \left\| \frac{d_\rho^{\pi^*}}{d_\mu^\pi} \right\|_\infty \max_{\pi_i' \in \Pi_i} \left\{ \sum_{s \in \mathcal{S}} d_\mu^\pi(s) \, \mathbb{E}_{a_i \sim \pi_i'(\cdot|s)} \mathbb{E}_{\mathbf{a}_{-i} \sim \pi_{-i}^*(\cdot|s)} \left[A_s^i(\pi, a_i, \mathbf{a}_{-i}) \right] \right\}.$$

To proceed, observe that

$$\mathbb{E}_{a_i \sim \pi_i(\cdot|s)} \mathbb{E}_{\mathbf{a}_{-i} \sim \pi_{-i}^*(\cdot|s)} \left[A_s^i(\pi, a_i, \mathbf{a}_{-i}) \right] = 0.$$

Thus, for any $\pi_i' \in \Pi_i$ and any state $s \in \mathcal{S}$, it holds that

$$\mathbb{E}_{a_i \sim \pi_i'(\cdot|s)} \mathbb{E}_{\mathbf{a}_{-i} \sim \pi_{-i}^*(\cdot|s)} \left[A_s^i(\pi, a_i, \mathbf{a}_{-i}) \right] =$$

$$= \sum_{a_i \in \mathcal{A}_i} \left(\pi_i'(a_i \mid s) - \pi_i(a_i \mid s) \right) \mathbb{E}_{\mathbf{a}_{-i} \sim \pi_{-i}^*(\cdot|s)} \left[A_s^i(\pi, a_i, \mathbf{a}_{-i}) \right]$$

$$= \sum_{a_i \in \mathcal{A}_i} \left(\pi_i'(a_i \mid s) - \pi_i(a_i \mid s) \right) \mathbb{E}_{\mathbf{a}_{-i} \sim \pi_{-i}^*(\cdot|s)} \left[Q_s^i(\pi, a_i, \mathbf{a}_{-i}) \right]$$

since $V_s^i(\pi)$ does not depend on a_i. Substituting back in the last inequality of the previous calculations, we obtain that

$$\Phi_\rho(\pi^*) - \Phi_\rho(\pi) \leq$$

$$\leq \left\| \frac{d_\rho^{\pi^*}}{d_\mu^\pi} \right\|_\infty \max_{\pi_i' \in \Pi_i} \left\{ \sum_{s,a_i} \frac{d_\mu^\pi(s)}{1-\gamma} \left(\pi_i'(a_i \mid s) - \pi_i(a_i \mid s) \right) \mathbb{E}_{\mathbf{a}_{-i} \sim \pi_{-i}^*(\cdot|s)} \left[Q_s^i(\pi, a_i, \mathbf{a}_{-i}) \right] \right\}$$

$$= \left\| \frac{d_\rho^{\pi^*}}{d_\mu^\pi} \right\|_\infty \max_{\pi_i' \in \Pi_i} (\pi_i' - \pi_i)^\top \nabla_{\pi_i} V_\mu^i(\pi),$$

where we used the policy gradient theorem ([1, 32]) under the assumption of direct policy parameterization (cf. equation (2)). We can further upper bound the last expression by using that $d_\mu^\pi(s) \geq (1-\gamma)\mu(s)$ which follows immediately from the definition of the discounted visitation distribution $d_\mu^\pi(s)$ for any initial state distribution μ. Finally, property P2 of Proposition B.1, implies that $\nabla_{\pi_i} V_\rho^i(\pi) = \nabla_{\pi_i} \Phi_\rho(\pi)$ (making crucial use of the MPG structure). Putting these together, we have that

$$\Phi_\rho(\pi^*) - \Phi_\rho(\pi) \leq \frac{1}{1-\gamma} \left\| \frac{d_\rho^{\pi^*}}{\mu} \right\|_\infty \max_{\pi'=(\pi_i', \pi_{-i}^*)} (\pi' - \pi)^\top \nabla_{\pi_i} V_\mu^i(\pi)$$

$$= \frac{1}{1-\gamma} \left\| \frac{d_\rho^{\pi^*}}{\mu} \right\|_\infty \max_{\pi'=(\pi_i', \pi_{-i}^*)} (\pi' - \pi)^\top \nabla_{\pi_i} \Phi_\mu(\pi),$$

as claimed. □

Proof of Lemma 4.4. It suffices to show that the maximum eigenvalue in absolute value of the Hessian of Φ is at most $\frac{2n\gamma A_{\max}}{(1-\gamma)^3}$, i.e., that

$$\left\|\nabla^2\Phi_\mu\right\|_2 \leq \frac{2n\gamma A_{\max}}{(1-\gamma)^3}.$$

We first prove the following intermediate claim.

Claim C.2. *Consider the symmetric block matrix C with $n \times n$ matrices so that $\|C_{ij}\|_2 \leq L$. Then, it holds that $\|C\|_2 \leq nL$, i.e., if all block submatrices have spectral norm at most L, then C has spectral norm at most nL.*

Proof. We will prove the claim by induction on n. For $n = 2$ we need to show that

$$\|C\|_2 := \left\|\begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix}\right\|_2 \leq 2L$$

if $\|C_{11}\|_2, \|C_{12}\|_2, \|C_{21}\|_2, \|C_{22}\|_2 \leq L$. Define matrix W to be

$$W := 2L \cdot I - C = \begin{pmatrix} 2L \cdot I - C_{11} & -C_{12} \\ -C_{21} & 2L \cdot I - C_{22} \end{pmatrix},$$

where I is the identity matrix (of appropriate size). If we show that W is positive semi-definite, then it follows that W has only non-negative eigenvalues, which, in turn, implies that the spectral norm of C is at most $2L$. To see this, set

$$W_1 := \begin{pmatrix} L \cdot I - C_{11} & 0 \\ 0 & L \cdot I - C_{22} \end{pmatrix}, W_2 := \begin{pmatrix} L \cdot I & -C_{12} \\ -C_{21} & L \cdot I \end{pmatrix}.$$

W_1 is positive semi-definite as a block diagonal matrix with diagonal blocks positive semi-definite matrices. Moreover, by Schur complement we get that W_2 is positive semi-definite as long as $L \cdot I$ is positive semi-definite and $L \cdot I - \frac{1}{L} \cdot C_{12}C_{21}$ is positive semidefinite. By assumption, we have that

$$\frac{1}{L}\|C_{12}C_{21}\|_2 \leq \frac{1}{L}\|C_{12}\|_2\|C_{12}\|_2 \leq L,$$

which implies that $L \cdot I - \frac{1}{L} \cdot C_{12}C_{21}$ has non-negative eigenvalues. Thus, W_2 is positive semi-definite. We conclude that $W_1 + W_2$ is positive semi-definite (sum of positive semi-definite matrices is positive semi-definite) and the claim follows.

For the induction step, suppose that the claim holds for an $n = k - 1 \geq 2$. To establish that it also holds for k, we need to show that

$$\|C\|_2 := \left\|\begin{pmatrix} C_{11} & C_{12} & \ldots & C_{1k} \\ C_{21} & C_{22} & \ldots & C_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ C_{k1} & C_{k2} & \ldots & C_{kk} \end{pmatrix}\right\|_2 \leq kL$$

34

as long as $\|C_{ij}\|_2 \leq L$ for all i, j. Let $W = kL \cdot I - C$. To show that W is positive semi-definite consider

$$W_1 := \begin{pmatrix} kL \cdot I - C_{11} & -C_{12} & -C_{13} & \ldots & -C_{1k} \\ -C_{21} & L \cdot I & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & \\ -C_{k1} & 0 & 0 & \ldots & L \cdot I \end{pmatrix}, W_2 := W - W_1.$$

By induction, it follows that W_2 is positive semi-definite. We need to show that the same holds for W_1. By Schur complement we obain that W_1 is positive semi-definite if and only if $kL \cdot I - C_{11} - \frac{1}{L} \sum_{i=2}^{k} C_{1i} C_{i1}$ is positive semi-definite. It follows that

$$\left\| C_{11} - \frac{1}{L} \sum_i C_{1i} C_{i1} \right\|_2 \leq \|C_{11}\|_2 + \frac{1}{L} \sum_{i=2}^{k} \|C_{1i}\|_2 \|C_{i1}\|_2 \leq L + (k-1)L = kL.$$

Hence W_1 is positive semi-definite and the induction is complete. $\qquad\square$

Returning to the statement of Lemma 4.4, we will show that

$$\left\| \nabla^2_{\pi_j \pi_i} V^j_\mu \right\|_2 \leq C, \tag{21}$$

for all $i, j \in \mathcal{N}$ with C chosen to be $\frac{2\gamma A_{\max}}{(1-\gamma)^3}$. Assuming we have shown (21), we conclude from Claim C.2 that

$$\left\| \nabla^2 \Phi_\mu \right\|_2 \leq nC,$$

and hence Φ will be nC-smooth (the proof of Lemma 4.4 will follow).

To prove (21), we follow the same proof steps as in the proof of [1], Lemma D.3. We will need to prove an upper bound on the largest eigenvalue (in absolute value) of the matrix

$$\nabla^2_{\pi_j \pi_i} V^j_\mu = \nabla^2_{\pi_j \pi_i} V^i_\mu,$$

along the direction where only agent i is allowed to change policy.

Fix policy $\pi = (\pi_1, ..., \pi_n)$, agents $i \neq j$, scalars $t, s \geq 0$, state s_0 and u, v be unit vectors such that $\pi_i + t \cdot u \in \Delta(\mathcal{A}_i)^S$ and $\pi_j + s \cdot v \in \Delta(\mathcal{A}_j)^S$. Moreover, let $V(t) = V^i_{s_0}(\pi_i + t \cdot u, \pi_{-i})$. and $W(t, s) = V^i_{s_0}(\pi_i + t \cdot u, \pi_j + s \cdot v, \pi_{-i,-j})$. It suffices to show that

$$\max_{\|u\|_2=1} \left| \frac{d^2 V(0)}{dt^2} \right| \leq \frac{2\gamma |\mathcal{A}_i|}{(1-\gamma)^3} \text{ and } \max_{\|u\|_2=1} \left| \frac{d^2 W(0,0)}{dtds} \right| \leq \frac{2\gamma \sqrt{|\mathcal{A}_i||\mathcal{A}_j|}}{(1-\gamma)^3}. \tag{22}$$

- We first focus on $V(t)$.

It holds that $V(t) = \sum_{a \in \mathcal{A}_i} \sum_{\mathbf{a} \in \mathcal{A}_{-i}} (x_{i,s_0,a} + tu_{i,s_0,a}) \prod_{j \neq i} x_{j,s_0,a_j} Q^i_{s_0}((\pi_i + tu, \pi_{-i}), (a, \mathbf{a}))$ (note that $\sum_{a \in \mathcal{A}_i} \sum_{\mathbf{a} \in \mathcal{A}_{-i}} (x_{i,s_0,a} + tu_{i,s_0,a}) \prod_{j \neq i} x_{j,s_0,a_j} = 1$ since it is a distribution), hence taking the second derivative we have

$$\begin{aligned} \frac{d^2 V(0)}{dt^2} &= \sum_{a \in \mathcal{A}_i} \sum_{\mathbf{a} \in \mathcal{A}_{-i}} (x_{i,s_0,a} + tu_{i,s_0,a}) \prod_{j \neq i} x_{j,s_0,a_j} \frac{d^2 Q^i_{s_0}(\pi, (a, \mathbf{a}))}{dt^2} \\ &\quad + 2 \sum_{a \in \mathcal{A}_i} \sum_{\mathbf{a} \in \mathcal{A}_{-i}} u_{i,s_0,a} \prod_{j \neq i} x_{j,s_0,a_j} \frac{d Q^i_{s_0}(\pi, (a, \mathbf{a}))}{dt} \end{aligned} \tag{23}$$

For the remaining of the first part of the proof, we shall show

$$\left|\frac{dQ^i_{s_0}(\pi,(a,\mathbf{a}))}{dt}\right| \leq \frac{\gamma\sqrt{|\mathcal{A}_i|}}{(1-\gamma)^2} \text{ and } \left|\frac{d^2Q^i_{s_0}(\pi,(a,\mathbf{a}))}{dt^2}\right| \leq \frac{2\gamma^2|\mathcal{A}_i|}{(1-\gamma)^3},$$

and then combining with (23) we get

$$\left|\frac{d^2V(0)}{dt^2}\right| \leq \frac{2\gamma\sqrt{|\mathcal{A}_i|}}{(1-\gamma)^2}\sum_{a\in\mathcal{A}_i}|u_{i,s_0,a}| + \frac{2\gamma^2|\mathcal{A}_i|}{(1-\gamma)^3}$$

$$\leq \frac{2\gamma|\mathcal{A}_i|}{(1-\gamma)^2} + \frac{2\gamma^2|\mathcal{A}_i|}{(1-\gamma)^3}$$

$$\leq \frac{2\gamma A_{\max}}{(1-\gamma)^3}.$$

To bound the derivative of the Q-function, observe that $Q^i_{s_0}((\pi_i + tu, \pi_{-i}), (a, \mathbf{a})) = e^\top_{s_0,a}(I - \gamma P(t))^{-1}r$, where $r(s_0, a)$ is the expected reward of agent i (w.r.t the randomness of the remaining agents) if he chooses action a at state s_0 and $P(t)$ is state-action transition matrix of w.r.t the joint distribution of all agents but i, i.e., π_{-i} and the environment.

It is clear that $\frac{d^2P}{dt^2} = 0$ (linear with respect to t because of direct parametrization) and moreover $\left\|\frac{dP}{dt}\right\|_\infty \leq \sum_{a\in\mathcal{A}_i}|u_{i,s_0,a}| \leq \sqrt{\mathcal{A}_i} \leq \sqrt{A_{\max}}$. Using the fact that $\|(I - \gamma P(t))^{-1}\|_\infty \leq \frac{1}{1-\gamma}$, we get

$$\left|\frac{dQ^i_{s_0}(\pi,(a,\mathbf{a}))}{dt}\right| = \gamma\left|e^\top_{s_0,a}(I - \gamma P(0))^{-1}\frac{dP(0)}{dt}(I - \gamma P(0))^{-1}r\right| \leq \frac{\gamma\sqrt{|\mathcal{A}_i|}}{(1-\gamma)^2} \leq \frac{\gamma\sqrt{A_{\max}}}{(1-\gamma)^2}, \quad (24)$$

and also

$$\left|\frac{d^2Q^i_{s_0}(\pi,(a,\mathbf{a}))}{dt^2}\right| = 2\gamma^2\left|e^\top_{s_0,a}(I - \gamma P(0))^{-1}\frac{dP(0)}{dt}(I - \gamma P(0))^{-1}\frac{dP(0)}{dt}(I - \gamma P(0))^{-1}r\right|$$

$$\leq \frac{2\gamma^2|\mathcal{A}_i|}{(1-\gamma)^3} \leq \frac{2\gamma^2 A_{\max}}{(1-\gamma)^3}, \quad (25)$$

Since u is arbitrary, the first part of (22) is proved.

- For the second part, we focus on $W(t)$ which is equal to

$$W(t,s) = \sum_{a\in\mathcal{A}_i}\sum_{b\in\mathcal{A}_j}\sum_{\mathbf{a}\in\mathcal{A}_{-i,-j}}(x_{i,s_0,a} + tu_{i,s_0,a})(x_{j,s_0,b} + sv_{j,s_0,b})\cdot$$

$$\cdot\prod_{j'\neq i,j}x_{j',s_0,a_{j'}}Q^i_{s_0}((\pi_i + tu, \pi_j + sv, \pi_{-i,-j}), (a, b, \mathbf{a})) \quad (26)$$

We consider the derivative of W (26) and we get

$$
\begin{aligned}
\frac{dW(0,0)}{dtds} &= \sum_{a\in\mathcal{A}_i}\sum_{b\in\mathcal{A}_j}\sum_{\mathbf{a}\in\mathcal{A}_{-i,-j}} u_{i,s_0,a}v_{j,s_0,b}\cdot\prod_{j'\neq i,j} x_{j',s_0,a_{j'}} Q^i_{s_0}(\pi,(a,b,\mathbf{a})) \\
&+ \sum_{a\in\mathcal{A}_i}\sum_{\mathbf{a}\in\mathcal{A}_{-i}} u_{i,s_0,a}\cdot\prod_{j'\neq i} x_{j',s_0,a_{j'}}\frac{dQ^i_{s_0}(\pi,(a,\mathbf{a}))}{dt} \\
&+ \sum_{b\in\mathcal{A}_j}\sum_{\mathbf{a}\in\mathcal{A}_{-j}} v_{j,s_0,b}\cdot\prod_{j'\neq j} x_{j',s_0,a_{j'}}\frac{dQ^i_{s_0}(\pi,(b,\mathbf{a}))}{dt}. \\
&+ \sum_{\mathbf{a}\in\mathcal{A}}\prod_{j'} x_{j',s_0,a_{j'}}\frac{d^2Q^i_{s_0}(\pi,\mathbf{a})}{dtds}.
\end{aligned}
\tag{27}
$$

The first term of the sum in absolute value is at most $\frac{\sqrt{|\mathcal{A}_i||\mathcal{A}_j|}}{1-\gamma}$ (assuming rewards lie in $[0,1]$.)
Moreover using (24) the second term of the sum in absolute value is bounded by $\frac{\gamma\sqrt{|\mathcal{A}_i|}\sqrt{|\mathcal{A}_i|}}{(1-\gamma)^2}$
and the third term by $\frac{\gamma\sqrt{|\mathcal{A}_j|}\sqrt{|\mathcal{A}_i|}}{(1-\gamma)^2}$. To bound the $\frac{d^2Q^i_{s_0}(\pi,\mathbf{a})}{dtds}$, the same approach works that we
used to prove (25) with the extra fact that the state-action transition matrix is $P(t,s)$ and
moreover the reward $r(s_0,a,b)$ is the expected reward of agent i (w.r.t the randomness of all
agents but i,j) if i chooses action a and j chooses b at state s_0.

Finally for the fourth term we get that

$$
\begin{aligned}
\left|\frac{d^2Q^i_{s_0}(\pi,\mathbf{a})}{dtds}\right| &\leq \gamma^2\left|e^\top_{s_0,a}(I-\gamma P(0,0))^{-1}\frac{dP(0,0)}{ds}(I-\gamma P(0,0))^{-1}\frac{dP(0,0)}{dt}(I-\gamma P(0,0))^{-1}r\right| + \\
&+ \gamma^2\left|e^\top_{s_0,a}(I-\gamma P(0,0))^{-1}\frac{dP(0,0)}{dt}(I-\gamma P(0,0))^{-1}\frac{dP(0,0)}{ds}(I-\gamma P(0,0))^{-1}r\right| + \\
&+ \gamma\left|e^\top_{s_0,a}(I-\gamma P(0,0))^{-1}\frac{d^2P(0,0)}{dt^2}(I-\gamma P(0,0))^{-1}r\right| \\
&\leq \frac{\gamma^2\sqrt{|\mathcal{A}_i||\mathcal{A}_j|}}{(1-\gamma)^3}+\frac{\gamma^2\sqrt{|\mathcal{A}_i||\mathcal{A}_j|}}{(1-\gamma)^3}+\frac{\gamma\sqrt{|\mathcal{A}_i||\mathcal{A}_j|}}{(1-\gamma)^2}\leq\frac{2\gamma\sqrt{|\mathcal{A}_i||\mathcal{A}_j|}}{(1-\gamma)^3}\leq\frac{2\gamma A_{\max}}{(1-\gamma)^3}.
\end{aligned}
$$

$\square$

D   Auxiliary Lemmas

Recall that $P_\mathcal{X}$ denotes the projection onto some set $\mathcal{X}$.

Lemma D.1 ([6], Lemma 3.6). *Let f be a β-smooth function[6] with convex domain $\mathcal{X}$. Let $x\in\mathcal{X}$, $x^+ = P_\mathcal{X}(x-\frac{1}{\beta}\nabla f(x))$ and $g_\mathcal{X}(x) = \beta(x-x^+)$. Then the following holds true:*

$$
f(x^+) - f(x) \leq -\frac{1}{2\beta}\|g_\mathcal{X}(x)\|^2_2.
$$

---

[6]Differentiable with ∇f to be β-Lipschitz.

Lemma D.2 ([1], Proposition B.1). *Let $f(\pi)$ be a β-smooth in $\pi \in \Delta(\mathcal{A})^S$. Define the gradient mapping*

$$G(\pi) = \beta \left(P_{\Delta(\mathcal{A})^S} \left(\pi + \frac{1}{\beta} \nabla_\pi f(\pi) \right) - \pi \right)$$

and the update rule for the projected gradient is $\pi' = \pi + \frac{1}{\beta} G(\pi)$. If $\|G(\pi)\|_2 \leq \epsilon$ then

$$\max_{\pi + \delta \in \Delta(\mathcal{A})^S, \|\delta\|_2 \leq 1} \delta^\top \nabla_\pi f(\pi') \leq 2\epsilon.$$

Lemma D.3. *Let $f(x)$ be a β-smooth in where x uses α-greedy parametrization. Define the gradient mapping*

$$G(x) = \beta \left(P_{\Delta(\mathcal{A})^S} \left(x + \frac{1}{\beta} \nabla f(x) \right) - x \right)$$

and the update rule for the projected gradient is $x' = x + \frac{1}{\beta} G(x)$. If $\|G(x)\|_2 \leq \epsilon$ then

$$\max_{\pi + \delta \in \Delta(\mathcal{A})^S, \|\delta\|_2 \leq 1} \delta^\top \nabla f(x') \leq 2\epsilon + \sqrt{n \alpha A_{\max}}.$$

Proof. It is a direct application of Lemma D.2 and the fact that Φ_μ is Lipschitz with parameter $\frac{\sqrt{n A_{\max}}}{(1-\gamma)^2}$ (this is Proposition 3 in page 23 of [11]). $\qquad\square$

Lemma D.4 (Unbiased with bounding variance [11]). *It holds that $\hat{\nabla}_{\pi_i}^{(t)}$ is unbiased estimator of $\nabla_{\pi_i} V^i$ for all i, that is*

$$\mathbb{E}_{\pi^{(t)}} \hat{\nabla}_{\pi_i}^{(t)} = \nabla_{\pi_i} V_\rho^i(\pi^{(t)}) \text{ for all } i.$$

Moreover for all agents i we get that (this is what the authors actually prove)

$$\mathbb{E}_{\pi^{(t)}} \left\| \hat{\nabla}_{\pi_i}^{(t)} \right\|_2^2 \leq \frac{24 A_{\max}^2}{\alpha(1-\gamma)^4}.$$

E   Additional Experiments

In this part, we provide a systematic analysis of variations of the experimental setting in Section 5.

Coordination beyond MPGs   As mentioned in Remark 4, we know that policy gradient converges also in other cooperative settings which may fail to be exact MPGs. To study this case, we modify our experiment from Section 5. The setting remains mostly the same, except now in the distancing state the weights are no longer reduced by the same constant factor, i.e., $w_k^{\text{dist}} = w_k^{\text{safe}} - c$ for all facilities k, but are instead reduced by differing (yet still sufficiently large) c_k's, i.e., $w_k^{\text{dist}} = w_k^{\text{safe}} - c_k$ for $k = A, B, C, D$.

This change is such that the MDP no longer satisfies condition C2 of Proposition 3.2. Since the transitions depend on the actions of the players, a similar reasoning as in Example 2 suggests that the game is not an MPG. However, it is still natural that cooperation is still desirable in this setting. In particular, if the c_k's for all $k = A, B, C, D$ are taken to be greater than the c of the MPG from Section 5, then the agents can be said to have even "stronger" incentive to cooperate. The results of running independent policy gradient on this variant are shown in Figure 5. Independent policy gradient requires more iterations to converge compared to the MPG setting, but still arrives at the same Nash policy.

Figure 5: Figures similar to Figure 4, except now with $c > c_A > c_B > c_C > c_D$ as described in the text. Independent policy gradient requires more iterations to converge compared to the MPG setting, but still arrives at the same Nash policy.

Figure 6: Convergence to deterministic Nash policies of independent policy gradient in a variation of the MPG of Section 5 with $N = 16$ agents and $A_i = 5$ facilities, $\mathcal{A}_i = \{A, B, C, D, E\}$ with $w_A < w_B < w_C < w_D < w_E$ (i.e., E is the most preferable by all agents). Again, while there are several (symmetric) deterministic Nash policies, all of them yield the same distribution of agents among states (leftmost panel). All runs converge successfully to that outcome (however, some runs required a larger number of iterates to converge).

Coordination with more agents and facilities We next test the performance of the independent policy gradient algorithm in a larger setting with $N = 16$ agents and $A_i = 5$ facilities, $\mathcal{A}_i = \{A, B, C, D, E\}$ with $w_A < w_B < w_C < w_D < w_E$ (i.e., E is the most preferable by all agents). We use a learning rate $\eta = 0.0001$ for all agents (which is again much larger than the theoretical guarantee of Theorem 4.5). All runs lead to convergence to an (optimal) Nash policy as shown in the middle and rightmost panels. The leftmost panel shows the distribution of the agents among facilities in both states, which is the same (and the optimal one) in all Nash policies that are reached by the algorithm. The results are shown in Figure 6.

Coordination with random transitions Next, we study the effect of adding randomness to the transitions on the performance of the individual policy gradient algorithm. In this case, we experiment with the same setting as in Section 5 (i.e., $N = 8$ agents and $A_i = 4$ facilities that each agent $i \in \mathcal{N}$ can choose from), but use the following stochastic transition rule instead: in addition to the existing transition rules, the sequence of play may transition from the safe to the distancing state with probability $p\%$ regardless of the distribution of the agents and may remain at the distancing state with probability $q\%$ again regardless of the distribution of the agents there.
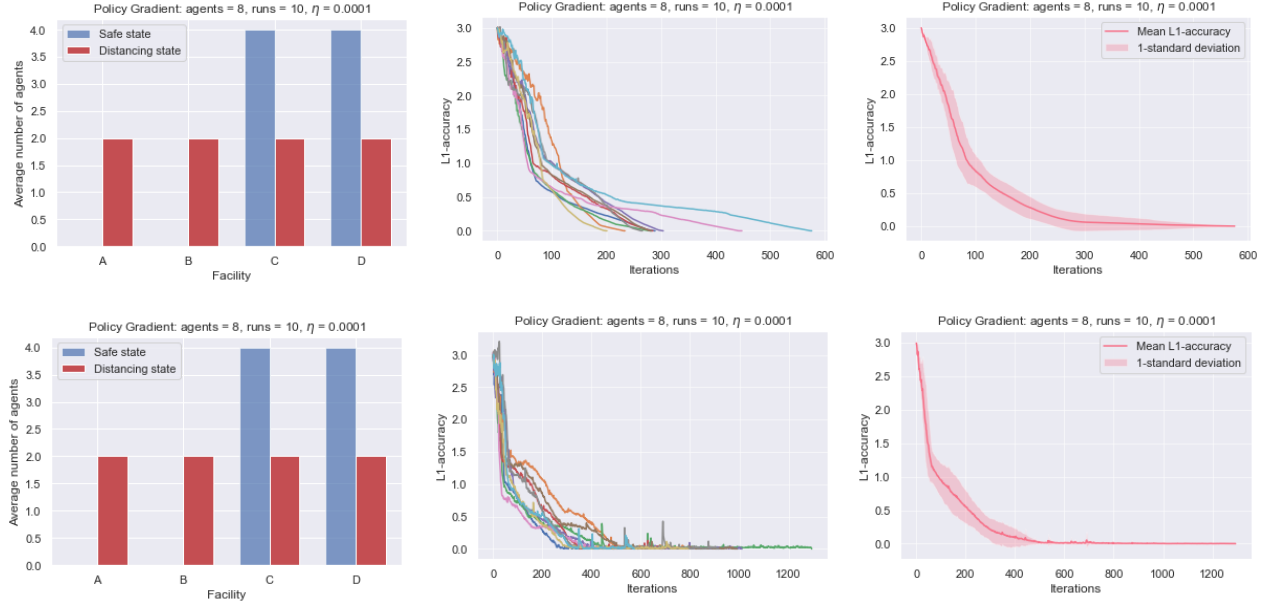
Figure 7: Convergence to deterministic Nash policies of independent policy gradient in two variations of the MPG of Section 5 with stochastic transitions between states.

Two sets of results are presented in Figure 7. In the first (upper panels), we use $p, q = 1\%, 10\%$ and in the second $p, q = 5\%, 20\%$. In both cases, we use a learning rate $\eta = 0.0001$ (several orders of magnitude higher than what is required by Theorem 4.7). Independent policy gradient converges in both cases to deterministic Nash policies despite the randomness in the transitions. However, for higher levels of randomness (lower panels), the algorithm remains at an ϵ-Nash policy for a high number of iterations. This is in line with the theoretical predictions of Theorem 4.7.