Algorithm 1 ALMANAC

Input: specifications $\{\varphi^j\}_{0\le j\le m}$, discount rates γ_V, γ_U, learning rates $\alpha, \beta^V, \beta^U, \eta, \iota$, reset probability p
Output: policy π^i_*

convert each φ^j into an LDBA B^j
initialise parameters θ^i, x^i, $\{v^j\}_{1\le j\le m}$, $\{u^j\}_{1\le j\le m}$, λ^i
while θ^i not converged **do**
  while x^i not converged **do**
    initialise $t \leftarrow 0$, $end \leftarrow \bot$, and $Z^j \leftarrow \varnothing$ for each j
    sample $s_0^\otimes \sim \zeta^\otimes$
    while $end = \bot$ **do**
      $Z^j \leftarrow Z \cup \{s_t^\otimes\}$ for each j
      sample $a_t^i \sim \pi_\theta^i(\cdot|s_t^\otimes)$
      observe $s_{t+1}^\otimes$ and r_{t+1}^j for each j
      if $r_{t+1}^j > 0$ **or** $\phi(s_{t+1}^\otimes)^\top v^j = 0$ **then**
        for $s_k^\otimes \in Z^j$ **do** update v^j using (5)
        $Z^j \leftarrow \varnothing$
      update u^j using (6) for each j
      update x^i and λ^i using (7)
      with probability p **set** $end \leftarrow \top$
  update θ^i using (8)
return π^i

4.3 Convergence and Correctness

By making use of results from the stochastic approximation and RL literature we provide an asymptotic convergence guarantee to locally or globally optimal joint policies with respect to multiple LTL specifications, depending on whether agents use local or global policies respectively. We assume that the following conditions hold:

(1) S and A are finite, and all reward functions are bounded.
(2) The Markov chain induced by any θ is irreducible over $S^\otimes$.
(3) $\pi^i(a^i|s^\otimes;\theta^i)$ is continuously differentiable $\forall i, s^\otimes, a^i$
(4) Let Φ be the $|S^\otimes| \times c$ matrix with rows $\phi(s^\otimes)$. Then Φ has full rank, $c \le |S|$, and $\nexists w \in W$ such that $\Phi w = 1$.
(5) $\mathbb{E}_t\left[L^i_{V^j}(x_t^{i*};\theta_t,v_*^j)\right] \le e^j_{approx}$, where e^j_{approx} is some constant, thus $\mathbb{E}_t\left[L^i_V(x_t^{i*};\theta_t,v_*)\right] \le e_{approx} \coloneqq \sum_j w[j] e^j_{approx}$.
(6) $\exists \sigma < \infty$ s.t. $\log \pi^i(a^i|s^\otimes;\theta)$ is a *σ-smooth* in θ^i $\forall i, s^\otimes, a^i$.
(7) The *relative condition number* is finite.
(8) $\pi^i(\cdot|s^\otimes;\theta^i)$ is initialised as the uniform distribution $\forall i, s^\otimes$.

Conditions 1–4 are standard within the literature on the convergence of actor critic algorithms [6, 26]. Conditions 5–8 are taken from recent work on the convergence of natural policy gradient methods by Agarwal et al. [1]. Of particular note is condition 5, where $e_{approx} = 0$ when π^i is a sufficiently rich class, such as an over-parametrised neural network. We recall that if $\log \pi^i(a^i|s^\otimes;\theta)$ is a *σ-smooth function* of θ^i then for any $\theta^i_1, \theta^i_2 \in \Theta^i$ we have:

$$\left\|\nabla_{\theta^i} \log \pi^i(a|s^\otimes;\theta^i_1) - \nabla_{\theta^i} \log \pi^i(a|s^\otimes;\theta^i_2)\right\|_2 \le \alpha \left\|\theta^i_1 - \theta^i_2\right\|_2.$$

Regarding 6 we define $\Sigma^\nu(\theta^i) \coloneqq \mathbb{E}_{(s^\otimes,a)\sim\nu}\left[\psi^i_{\theta^i}(a^i_t|s^\otimes_t)\psi^i_{\theta^i}(a^i_t|s^\otimes_t)^\top\right]$ where ν is some state-action distribution. Then the average *relative condition number* [1] is defined and bounded as follows for each player i and each specification φ^j:

$$\mathbb{E}\left[\sup_{x^i} \frac{{x^i}^\top \Sigma_{\nu^j_*}(\theta^i_t) x^i}{{x^i}^\top \Sigma_\xi(\theta^i_t) x^i}\right] \le \kappa,$$

where ξ is some initial state-action distribution and:

$$\nu^j_* \coloneqq \nu^j_{\theta,\xi}(s^\otimes, a) = \sum_{(s_0^\otimes,a_0)\in S^\otimes\times A} \xi^\otimes(s_0^\otimes, a_0) \sum_\rho \Pr^\theta_{G_B}(\rho|s_0^\otimes, a_0) \cdot \left[\frac{1}{\sum_{t=0}^\infty \Gamma^j_{0:t}} \sum_{t=0}^\infty \Gamma^j_{0:t} \mathbb{I}\big(\rho[t,t+0.5] = (s^\otimes, a)\big)\right]$$

and $\rho[t+0.5]$ refers to the action taken along the trajectory ρ at time t. Due to space limitations we refer the interested reader to the cited works above for further discussion of these conditions.

Our proof follows the recent work of Agarwal et al. [1]. We begin with a variant of the well-known performance difference lemma [25], using which we prove an analogue of the 'no regret' lemma from Agarwal et al. which is in turn based on the mirror-descent approach of Even-Dar et al. [14]. The proofs are similar to the originals, and so we relegate them to the supplementary material.

Lemma 2. *Suppose that $V_\theta(s^\otimes) \ge V_{\theta'}(s^\otimes)$ for some state $s^\otimes$ and two policies π and π' parametrised by θ and θ' respectively. Then:*

$$V_\theta(s^\otimes) - V_{\theta'}(s^\otimes) \le \sum_j w[j]\left(\mathbb{E}_\rho\left[\sum_{t=0}^\infty \Gamma^j_{0:t} A^j_{\theta'}(s_t^\otimes, a_t) \,\middle|\, F^j(\rho) = \infty\right]\right).$$

Lemma 3. *Consider a sequence of natural gradient updates $\{x^i_t\}_{0\le t\le T}$ found by ALMANAC such that $\|x^i_t\|_2 \le X$ for all t. Let us write $\iota_{0:T} = \sum_{t=0}^T \iota_t$, and recall that $F^j(\rho)$ is the number of times a path ρ in G_B passes through the accepting set F^j of automaton B^j. Let us write $\mathbb{E}_{\rho^*}$ instead of $\mathbb{E}_{\rho\sim \Pr^{\theta_*}_{G_B}(\cdot|s^\otimes), s^\otimes\sim\zeta^\otimes}$ and define e^j_t by:*

$$e^j_t \coloneqq \mathbb{E}_{\rho^*}\left[\sum_{\tau=0}^\infty \Gamma^j_{0:\tau}\left(A^j_{\theta_t}(s^\otimes_\tau, a_\tau) - \psi^i_{\theta^i_t}(a^i_\tau|s^\otimes_\tau)^\top x^i_t\right) \,\middle|\, F^j(\rho) = \infty\right],$$

where τ indexes ρ, i.e., $\rho[\tau] = s^\otimes_\tau$. Then we have:

$$V_{\theta_*}(s^\otimes) - \lim_{T\to\infty} \mathbb{E}_{t\sim\iota_T}\left[V_{\theta_t}(s^\otimes)\right] = \lim_{T\to\infty} \mathbb{E}_{t\sim\iota_T}\left[\sum_j w[j] e^j_t\right],$$

where we define the distribution ι_T over t with $\iota_T(t) \coloneqq \frac{\iota_t}{\iota_{0:T}}$.

Finally, we use these results to prove that ALMANAC converges to either locally or globally optimal joint policies (i.e., either an SPE or a team-optimal SPE in the original MG) depending on whether agents use local or global policy parameters. By *local* policy parameters we mean that the parameters θ^i stored and updated by agent i only define π^i, and thus $\pi(a|s^\otimes;\theta) = \prod_i \pi^i(a^i|s^\otimes;\theta^i)$ is limited in its representational power due to its factorisation. If, instead, agents share a random seed and each $\theta^i = \theta$ is sufficient to parametrise the whole joint policy π (hence *global*) then at each timestep every agent i can sample the same full joint action $a = (a^1, \ldots, a^n)$ and simply perform its own action a^i. As rewards are shared between agents then this means that updates to each agent's version of v, u, and x^i will also be identical, and therefore so too will updates to $\theta^i = \theta$. Though more expensive in terms of computation and memory, the use of global parameters guarantees convergence to the globally optimal joint policy.

Theorem 1. *Given an MG G and LTL objectives $\{\varphi^j\}_{1\leq j\leq m}$ (each equivalent to an LDBA B^j), let $G_B = G \otimes B^1 \otimes \cdots \otimes B^m$ be the resulting product MG with newly defined reward functions $R^j_\otimes$ and state-dependent discount functions Γ^j. Assume that γ_V satisfies Proposition 1, that the learning rates $\alpha, \beta^V, \beta^U, \eta, \iota$ are as in (9) and that conditions 1–8 hold. Then if each agent i uses* local *(*global*) parameters θ^i with* local *policy $\pi^i_{\theta^i}$ (*global *policy $\pi^i_{\theta^i} = \pi_\theta$) then as $T \to \infty$,* Almanac *converges to within*

$$\lim_{T\to\infty} \mathbb{E}_{t\sim\iota_T}\left[\sum_j w[j]\sqrt{e^j_{approx}}\frac{M^j}{(1-\gamma_V)P^j}\right]$$

of a local *(*global*) optimum of $\sum_j w[j]\Pr^\pi_G(s \models \varphi^j)$, where P^j and M^j are constants.*

Proof (Sketch). The proof proceeds via a multi-timescale stochastic approximation analysis and is asymptotic in nature [7]. We consider convergence of the critics, natural gradients, and actor in three steps, dividing our attention between the local and global settings, where required. **Step 1.** The convergence proof for the critics follows that of Tsitsiklis and Van Roy [54]. The hasty critic recursion is simply the classic linear semi-gradient temporal difference algorithm [49] which is known to converge to the unique TD fixed point with probability 1. A similar argument can be made for the patient critic. By waiting to update v^j until seeing a reward, we ensure that a discount is applied and thus that the patient critic recursion forms a contraction. The proof follows immediately from previous work [54], but using a k-step version of the relevant Bellman equation. **Step 2.** Due to the learning rates chosen according to (9) we may consider the more slowly updated parameters fixed for the purposes of analysing the convergence of more quickly updated parameters [7]. As the critic updates fastest we may consider it converged, and since the policy is only updated in the outer loop then it is fixed with respect to the natural gradient and Lagrange multiplier updates. We show that these updates form unbiased estimates of the relevant gradients and thus discrete approximations of the following ODEs:

$$\dot{x^i_t} = \Omega_{x^i}\left[-\nabla_{x^i} L^i_V(x^i;\theta,\nu)\right]$$
$$\dot{x^i_t} = \Omega_{x^i}\Big[-\nabla_{x^i}\big(L^i_U(x^i;\theta,\mu) + \lambda^i\big(L^i_V(x^i;\theta,\nu) - l^i\big)\big)\Big],$$
$$\dot{\lambda^i_\tau} = \Omega_{\lambda^i}\Big[\nabla_{\lambda^i}\big(L^i_U(x^i(\lambda^i_\tau);\theta,\mu) + \lambda^i\big[L^i_V(x^i(\lambda^i_\tau);\theta,\nu) - l^i\big)\big)\Big],$$

on timescales β^V, β^U, and η, respectively. Due to the convexity of L^i_V and L^i_U it can be shown that the recursions above lexicographically miminise L^i_V and then L^i_U and hence that the gradient x^i_* satisfies (3) [44]. **Step 3.** Finally we use Lemma 3 and bound each term e^j_t by $\sqrt{e^j_{approx}}\frac{M^j}{(1-\gamma_V)P^j}$ where M^j and P_j are constants. In particular, we have: $M^j := \max_k \mathbb{E}_{\rho^{\zeta}_*}\left[M^j_\rho(k) \,\middle|\, F^j(\rho) = \infty\right]$ where $M^j_\rho(k)$ is the number of steps along trajectory ρ between the k^{th} reward and preceding reward, and $P^j := \min\left(\sum_\rho \Pr^{\theta_*}_{G_B}\mathbb{I}(F^j(\rho) = \infty), 1\right)$. The proof structure follows that of Agarwal et al. [1] with minor variations to handle our use of multiple agents and multiple state-dependent discount rates. □

	2^1	2^2	2^3	2^4	2^5	2^6	2^7	2^8	2^9	2^{10}
1	0.13	0.20	0.16	0.21	0.14	0.13	0.17	0.22	0.19	–
2	0.54	0.26	0.19	0.12	0.30	0.19	0.36	0.29	0.38	–
3	0.48	0.23	0.25	0.10	0.20	0.15	0.10	0.31	–	–
4	0.44	0.21	0.02	0.16	0.22	0.26	0.23	0.34	–	–
5	0.17	0.30	0.10	0.13	0.22	0.06	0.30	–	–	–
1	0.14	0.07	0.11	0.17	0.18	0.09	0.24	0.14	0.25	–
2	0.15	0.07	0.15	0.34	0.20	0.15	0.17	0.06	–	–
3	0.15	0.14	0.12	0.25	0.23	0.52	0.28	–	–	–
4	0.12	0.25	0.23	0.23	0.22	–	–	–	–	–
5	0.23	0.21	0.28	0.45	0.01	–	–	–	–	–

Table 1: Average errors across a number of states (columns), agents (rows), and specifications (top and bottom).

5 EXPERIMENTS

Evaluating our proposed algorithm is non-trivial for several reasons. The first is its novelty; it is designed specifically to satisfy the non-Markovian, infinite-horizon specifications that other MARL algorithms are unable to learn, making a direct comparison less meaningful. The second is that the satisfaction of the specifications we wish to evaluate our algorithm against cannot be estimated simply from samples. For example, ψ may be true at every state in a set of samples despite $\mathrm{G}\,\psi$ being false with probability 1. Using a probabilistic model-checker instead raises a third and final difficulty, as even state-of-the-art tools are unable to handle the size of games or number of specifications that Almanac is applicable to.

Despite this, we provide an initial set of results in which we benchmark an implementation[2] of our algorithm against ground-truth models exported to PRISM, a probabilistic model-checker [27]. These results serve to demonstrate Almanac's empirical convergence properties, and how this performance varies as a function of the size of the state space, the number of actors, and the number of specifications (though, unfortunately, PRISM only supports multi-objective synthesis with two specifications). For each of these combinations, we randomly generated ten MGs and sample the specifications and weights. We then ran our algorithm for 5000 episodes and exported the resulting policy, game structure, and specifications to PRISM. The differences between the weighted sum of satisfaction probabilities resulting from Almanac and the ground-truth optimal quantities are displayed in Table 1. We ran PRISM with a maximum of 16GB of memory, 100,000 value iteration steps, and twelve hours of computation, but for some combinations this was insufficient.

ACKNOWLEDGMENTS

The authors thank Hosein Hasanbeig, Joar Skalse, Alper Kamil Bozkurt, Kaiqing Zhang, Salomon Sickert, and the anonymous reviewers for helpful comments. Hammond acknowledges the support of an EPSRC Doctoral Training Partnership studentship (Reference: 2218880) and the University of Oxford ARC facility.[3] Wooldridge and Abate acknowledge the support of the Alan Turing Institute.

[2] Our code can be found online at `https://github.com/lrhammond/almanac`.
[3] Details available at `http://dx.doi.org/10.5281/zenodo.22558`.

REFERENCES

[1] Alekh Agarwal, Sham M Kakade, Jason D Lee, and Gaurav Mahajan. 2020. Optimality and Approximation with Policy Gradient Methods in Markov Decision Processes. In *Proceedings of Thirty Third Conference on Learning Theory (Proceedings of Machine Learning Research, Vol. 125)*, Jacob Abernethy and Shivani Agarwal (Eds.). PMLR, 64–66.

[2] Shun'ichi Amari. 1998. Natural Gradient Works Efficiently in Learning. *Neural Computation* 10, 2 (1998), 251–276.

[3] Gurdal Arslan and Serdar Yuksel. 2017. Decentralized Q-learning for Stochastic Teams and Games. *IEEE Trans. Automat. Control* 62, 4 (2017), 1545–1558.

[4] Pranav Ashok, Jan Křetínský, and Maximilian Weininger. 2019. Pac Statistical Model Checking for Markov Decision Processes and Stochastic Games. In *Computer Aided Verification*. Springer International Publishing, 497–519.

[5] Dimitri P. Bertsekas and John N. Tsitsiklis. 1996. *Neuro-dynamic Programming*. Athena Scientific.

[6] Shalabh Bhatnagar, Richard S. Sutton, Mohammad Ghavamzadeh, and Mark Lee. 2009. Natural Actor–critic Algorithms. *Automatica* 45, 11 (2009), 2471–2482.

[7] Vivek S. Borkar. 2008. *Stochastic Approximation*. Hindustan Book Agency.

[8] Michael Bowling and Manuela Veloso. 2001. Rational and Convergent Learning in Stochastic Games. In *Proceedings of the 17th International Joint Conference on Artificial Intelligence - Volume 2* (Seattle, WA, USA) *(IJCAI'01)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1021–1026.

[9] Alper Kamil Bozkurt, Yu Wang, Michael M. Zavlanos, and Miroslav Pajic. 2019. Control Synthesis from Linear Temporal Logic Specifications Using Model-free Reinforcement Learning. *arXiv:1909.07299* (2019).

[10] Tomáš Brázdil, Krishnendu Chatterjee, Martin Chmelík, Vojtěch Forejt, Jan Křetiínský, Marta Kwiatkowska, David Parker, and Mateusz Ujma. 2014. Verification of Markov Decision Processes Using Learning Algorithms. In *Automated Technology for Verification and Analysis*. Springer International Publishing, 98–114.

[11] Lucian Busoniu, Robert Babuska, and Bart De Schutter. 2008. A Comprehensive Survey of Multiagent Reinforcement Learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 38, 2 (2008), 156–172.

[12] Vincent Conitzer and Tuomas Sandholm. 2003. Awesome: A General Multiagent Learning Algorithm That Converges in Self-play and Learns a Best Response against Stationary Opponents. In *Proceedings of the Twentieth International Conference on International Conference on Machine Learning (ICML'03)*. AAAI Press, Washington, DC, USA, 83–90.

[13] Kousha Etessami, Marta Kwiatkowska, Moshe Y. Vardi, and Mihalis Yannakakis. 2007. Multi-objective Model Checking of Markov Decision Processes. In *Tools and Algorithms for the Construction and Analysis of Systems*. Springer Berlin Heidelberg, 50–65.

[14] Eyal Even-Dar, Sham. M. Kakade, and Yishay Mansour. 2009. Online Markov Decision Processes. *Mathematics of Operations Research* 34, 3 (2009), 726–736.

[15] Dana Fisman, Orna Kupferman, and Yoad Lustig. 2010. Rational Synthesis. In *Tools and Algorithms for the Construction and Analysis of Systems*. Springer Berlin Heidelberg, 190–204.

[16] Jakob N. Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. 2018. Counterfactual Multi-agent Policy Gradients. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, Sheila A. McIlraith and Kilian Q. Weinberger (Eds.). AAAI Press, 2974–2982.

[17] Jie Fu and Ufuk Topcu. 2014. Probably Approximately Correct Mdp Learning and Control with Temporal Logic Constraints. In *Robotics: Science and Systems X, University of California, Berkeley, USA, July 12-16, 2014*, Dieter Fox, Lydia E. Kavraki, and Hanna Kurniawati (Eds.).

[18] Drew Fudenberg and Jean Tirole. 1991. *Game Theory*. The MIT Press.

[19] Ernst M. Hahn, Mateo Perez, Sven Schewe, Fabio Somenzi, Ashutosh Trivedi, and Dominik Wojtczak. 2019. Omega-regular Objectives in Model-free Reinforcement Learning. In *Tools and Algorithms for the Construction and Analysis of Systems*. Springer International Publishing, 395–412.

[20] Ernst M. Hahn, Mateo Perez, Sven Schewe, Fabio Somenzi, Ashutosh Trivedi, and Dominik Wojtczak. 2020. Reward Shaping for Reinforcement Learning with Omega-regular Objectives. *arXiv:2001.05977* (2020).

[21] Mohammadhosein Hasanbeig, Alessandro Abate, and Daniel Kroening. 2019. Certified Reinforcement Learning with Logic Guidance. *arXiv:1902.00778* (2019).

[22] Mohammadhosein Hasanbeig, Daniel Kroening, and Alessandro Abate. 2020. Deep Reinforcement Learning with Temporal Logics. In *Lecture Notes in Computer Science*. Springer International Publishing, 1–22.

[23] Kishor Jothimurugan, Rajeev Alur, and Osbert Bastani. 2019. A Composable Specification Language for Reinforcement Learning Tasks. In *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d Alché-Buc, E. Fox, and R. Garnett (Eds.). Curran Associates, Inc., 13041–13051.

[24] Sham Kakade. 2001. A Natural Policy Gradient. In *Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic* (Vancouver, British Columbia, Canada) *(NIPS'01)*. MIT Press, Cambridge, MA, USA, 1531–1538.

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

[26] Vijay R. Konda and John N. Tsitsiklis. 2000. Actor-critic Algorithms. In *Advances in Neural Information Processing Systems 12*, S. A. Solla, T. K. Leen, and K. Müller (Eds.). MIT Press, 1008–1014.

[27] Marta Kwiatkowska, Gethin Norman, and David Parker. 2011. Prism 4.0: Verification of Probabilistic Real-time Systems. In *Proc. 23rd International Conference on Computer Aided Verification (CAV'11) (LNCS, Vol. 6806)*, G. Gopalakrishnan and S. Qadeer (Eds.). Springer, 585–591.

[28] Marta Kwiatkowska, Gethin Norman, David Parker, and Gabriel Santos. 2019. Equilibria-based Probabilistic Model Checking for Concurrent Stochastic Games. In *Lecture Notes in Computer Science*. Springer International Publishing, 298–315.

[29] Borja G. León and Francesco Belardinelli. 2020. Extended Markov Games to Learn Multiple Tasks in Multi-Agent Reinforcement Learning. In *ECAI 2020 - 24th European Conference on Artificial Intelligence, 29 August-8 September 2020, Santiago de Compostela, Spain, August 29 - September 8, 2020 - Including 10th Conference on Prestigious Applications of Artificial Intelligence (PAIS 2020) (Frontiers in Artificial Intelligence and Applications, Vol. 325)*, Giuseppe De Giacomo, Alejandro Catalá, Bistra Dilkina, Michela Milano, Senén Barro, Alberto Bugarín, and Jérôme Lang (Eds.). IOS Press, 139–146.

[30] Xiao Li, Yao Ma, and Calin Belta. 2018. Automata Guided Reinforcement Learning with Demonstrations. *arXiv:1809.06305* (2018).

[31] Xiao Li, Cristian-Ioan Vasile, and Calin Belta. 2017. Reinforcement Learning with Temporal Logic Rewards. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE.

[32] Michael L. Littman. 1994. Markov Games As a Framework for Multi-agent Reinforcement Learning. In *Proceedings of the Eleventh International Conference on International Conference on Machine Learning* (New Brunswick, NJ, USA) *(ICML'94)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 157–163.

[33] Michael L. Littman, Ufuk Topcu, Jie Fu, Charles Isbell, Min Wen, and James MacGlashan. 2017. Environment-independent Task Specifications Via Gltl. *arXiv:1704.04341* (2017).

[34] Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, Pieter Abbeel, and Igor Mordatch. 2017. Multi-agent Actor-critic for Mixed Cooperative-competitive Environments. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Long Beach, California, USA) *(NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 6382–6393.

[35] Eric Maskin and Jean Tirole. 2001. Markov Perfect Equilibrium. *Journal of Economic Theory* 100, 2 (2001), 191–219.

[36] Chris Nota and Philip S. Thomas. 2020. Is the Policy Gradient a Gradient?. In *Proceedings of the 19th International Conference on Autonomous Agents and Multi-Agent Systems* (Auckland, New Zealand) *(AAMAS '20)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 939–947.

[37] Ann Nowé, Peter Vrancx, and Yann-Michaël De Hauwere. 2012. Game Theory and Multi-agent Reinforcement Learning. In *Adaptation, Learning, and Optimization*. Springer Berlin Heidelberg, 441–470.

[38] Ryohei Oura, Ami Sakakibara, and Toshimitsu Ushio. 2020. Reinforcement Learning of Control Policy for Linear Temporal Logic Specifications Using Limit-deterministic Generalized Büchi Automata. *arXiv:2001.04669* (2020).

[39] Julien Perolat, Bilal Piot, and Olivier Pietquin. 2018. Actor-critic Fictitious Play in Simultaneous Move Multistage Games. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics (Proceedings of Machine Learning Research, Vol. 84)*, Amos Storkey and Fernando Perez-Cruz (Eds.). PMLR, Playa Blanca, Lanzarote, Canary Islands, 919–928.

[40] Jan Peters and Stefan Schaal. 2008. Natural Actor-critic. *Neurocomputing* 71, 7-9 (2008), 1180–1190.

[41] Amir Pnueli. 1977. The Temporal Logic of Programs. In *Proceedings of the 18th Annual Symposium on Foundations of Computer Science (FOCS '77)*. IEEE Computer Society, USA, 46–57.

[42] H.L. Prasad, Prashanth L.A., and Shalabh Bhatnagar. 2015. In *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems* (Istanbul, Turkey) *(AAMAS '15)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 1371–1379.

[43] Guannan Qu, Adam Wierman, and Na Li. 2020. Scalable Reinforcement Learning of Localized Policies for Multi-agent Networked Systems. *arXiv:1912.02906* (2020).

[44] Mark J. Rentmeesters, Wei K. Tsai, and Kwei-Jay Lin. 1996. A Theory of Lexicographic Multi-criteria Optimization. In *Proceedings of ICECCS 1996: 2nd IEEE International Conference on Engineering of Complex Computer Systems*. IEEE Comput. Soc. Press.

[45] Dorsa Sadigh, Eric S. Kim, Samuel Coogan, S. Shankar Sastry, and Sanjit A. Seshia. 2014. A Learning Based Approach to Control Synthesis of Markov Decision Processes for Linear Temporal Logic Specifications. In *53rd IEEE Conference on Decision and Control, CDC 2014, Los Angeles, CA, USA, December 15-17, 2014*. 1091–1096.

[46] Salomon Sickert, Javier Esparza, Stefan Jaax, and Jan Křetínský. 2016. Limit-deterministic Büchi Automata for Linear Temporal Logic. In *Computer Aided Verification*. Springer International Publishing, 312–332.

[47] Joar Skalse, Lewis Hammond, and Alessandro Abate. 2021. Lexicographic Multi-objective Reinforcement Learning. *Forthcoming* (2021).

[48] Richard S. Sutton. 1988. Learning to Predict by the Methods of Temporal Differences. *Machine Learning* 3, 1 (1988), 9–44.

[49] Richard S. Sutton and Andrew G. Barto. 2018. *Reinforcement Learning*. The MIT Press.

[50] Richard S. Sutton, David McAllester, Satinder Singh, and Yishay Mansour. 1999. Policy Gradient Methods for Reinforcement Learning with Function Approximation. In *Proceedings of the 12th International Conference on Neural Information Processing Systems* (Denver, CO) *(NIPS'99)*. MIT Press, Cambridge, MA, USA, 1057–1063.

[51] Philip S. Thomas. 2014. Bias in Natural Actor-critic Algorithms. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32 (ICML'14)*. JMLR.org, Beijing, China, I–441–I–448.

[52] Rodrigo Toro Icarte, Toryn Klassen, Richard Valenzano, and Sheila McIlraith. 2018. Using Reward Machines for High-level Task Specification and Decomposition in Reinforcement Learning. In *Proceedings of the 35th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 80)*, Jennifer Dy and Andreas Krause (Eds.). PMLR, Stockholmsmässan, Stockholm Sweden, 2107–2116.

[53] Rodrigo Toro Icarte, Toryn Q. Klassen, Richard Valenzano, and Sheila A. McIlraith. 2018. Teaching Multiple Tasks to an Rl Agent Using Ltl. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems* (Stockholm, Sweden) *(AAMAS '18)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 452–461.

[54] John N. Tsitsiklis and Benjamin Van Roy. 1997. An Analysis of Temporal-difference Learning with Function Approximation. *IEEE Trans. Automat. Control* 42, 5 (1997), 674–690.

[55] Xiaofeng Wang and Tuomas Sandholm. 2002. 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)*. MIT Press, Cambridge, MA, USA, 1603–1610.

[56] Min Wen and Ufuk Topcu. 2016. Probably Approximately Correct Learning in Stochastic Games with Temporal Logic Specifications. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI'16)*. AAAI Press, New York, New York, USA, 3630–3636.

[57] Michael Wooldridge, Julian Gutierrez, Paul Harrenstein, Enrico Marchioni, Giuseppe Perelli, and Alexis Toumi. 2016. Rational Verification: From Model Checking to Equilibrium Checking. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI'16)*. AAAI Press, Phoenix, Arizona, 4184–4190.

[58] Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. 2019. Multi-agent Reinforcement Learning: A Selective Overview of Theories and Algorithms. *arXiv:1911.10635* (2019).

[59] Kaiqing Zhang, Zhuoran Yang, Han Liu, Tong Zhang, and Tamer Basar. 2018. Fully Decentralized Multi-agent Reinforcement Learning with Networked Agents. In *Proceedings of the 35th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 80)*, Jennifer Dy and Andreas Krause (Eds.). PMLR, Stockholmsmässan, Stockholm Sweden, 5872–5881.

[60] Martin Zinkevich, Amy Greenwald, and Michael L. Littman. 2005. Cyclic Equilibria in Markov Games. In *Proceedings of the 18th International Conference on Neural Information Processing Systems* (Vancouver, British Columbia, Canada) *(NIPS'05)*. MIT Press, Cambridge, MA, USA, 1641–1648.