Reducing the Global Carbon Footprint based on Multi-Agent Reinforcement Learning (MARL)

Valentin Kahn VALENTIN.KAHN@GMAIL.COM
Research Fellow, The School of AI

November 29, 2018

Abstract

This research paper intends to model the investment of groups of countries into carbon emission reductions based on a Mixed Markov Game setting, applying the principles of off-policy single-agent Reinforcement Learning to a multi-agent setting with a Markov Decision Process (MDP).
The study shows that countries which are choosing their carbon emission reduction actions under the constraint of optimizing their and their partners' mid-term economic benefit, achieve both higher cumulative rewards and higher reductions in their per capita CO_2 consumption, than their counterparts. It also shows that action choices considering immediate and future rewards for the individual agents, as well as the cumulative reward of all agents, converge to large reductions in the carbon emission level per capita of these countries. Multi-agent reinforcement learning (MARL) bears important problem-solving potential by modelling economic and political decision makers in simulated environments.

Keywords: reinforcement learning, game theory, climate change, global warming, global carbon emissions, global carbon footprint, multi-agent reinforcement learning, mixed markov games, markov decision process, q-learning, correlated equilibrium, temporal-difference learning

1 Climate Change and Game Theory

The following chapter introduces evidence and issues of climate change. The chapter talks about root causes, potential solutions, as well as economics of climate change and global warming. It finally models climate change as a non-cooperative game with an inefficient Nash equilibrium and sets the stage for multi-agent reinforcement learning (MARL) as an approach to overcoming this Nash equilibrium.

1.1 Evidence of Global Warming

This paper is not discussing the impact and consequences of global warming in detail and assumes that undertaking measures to preventing global warming is reasonable. According to Budd (2016), there are at least five indicators that the global climate is changing: the rise of the Earth's temperature, the loss of Arctic sea ice, the increase in mean sea level over the last 100 years, the increase in the number of events of extreme rainfall and the increased level of carbon dioxide in the atmosphere.

1.2 Root Causes of Global Warming

The reasons for man-made global warming lie in the well-known and -discussed greenhouse effect. So-called "heat-trapping gases" or "greenhouse gases" prevent the reflected thermal energy of the sun from escaping into space. While the existence of these gases is essential to sustain human life on earth, their abundance is a cause of global warming. And human activity has indeed led to increased concentration of heat-trapping gases in the atmosphere. Additionally, human-induced changes have reduced nature's capacity to absorb these gases, for instance through deforestation (Kelman, 2015). The main heat-trapping gases are carbon dioxide (CO_2), methane (CH_4), nitrous oxide (N_2O) and water vapor (H_2O, UCS USA, 2017). According to the Intergovernmental Panel on Climate Change IPCC (2013), of all these gases, carbon dioxide both has contributed by far the most to climate change between 1750 and 2011 and by far resides the longest in the atmosphere. Global CO_2 emissions are currently (2014) at 4.97 metric tons per capita, up from 3.1 metric tons in 1960. Additionally, the world's population has grown substantially over said time period (The World Bank, 2018). This paper assumes the consumption of CO_2 to be the main factor when it comes to human-made global warming and climate change and subsequently focuses on the reduction of said consumption. The "World Economic and Social Survey" by the United Nations in 2011 proposes to cap the CO_2 emissions per capita at 3 tons until 2050 (United Nations, 2011). According to the Intergovernmental Panel on Climate Change (IPCC, 2018), carbon emission intensity needs to be drastically reduced in order for us to be able to cap global warming to 1.5-2°C until 2100, as set by the Paris Agreement (United Nations Framework Convention on Climate Change (UNFCCC), 2018).

The Carbon Disclosure Project (CDP, 2017) released a report that attributes half of global industrial greenhouse gas emissions since 1751 (around 1 trillion tons of carbon dioxide) to not more than 100 fossil fuel extracting firms, most of them in the coal, oil or gas sectors. As these companies are bound to operate at locations and jurisdictions that contain large volumes of carbon-based natural resources, limiting or altering their operations (e.g. through large-scale abatement techniques, such as carbon capture and sequestration) and thus their emissions would be subject to regulations in these jurisdictions. However, the major driving force behind greenhouse gas emissions is not the production of fossil fuels, but the consumption thereof. The major consumers of fossil fuels and their products are large industrial and emerging economies, with China and the United States being responsible for about half of the world's greenhouse gas emissions (Talbot, 2014). In order to counteract global warming, it is the responsibility of these economies to reduce consumption of fossil fuels, by either leveraging alternative energy sources (such as renewable energies) and/or reducing their economic output overall.

1.3 Economics of Climate Change Mitigation

At this point in time, costs and benefits of reducing fossil fuel consumption cannot be estimated accurately. As emitted CO_2 lasts about 100 years in the atmosphere, costly reduction measures with high uncertainty about timing and size of incurred benefits in the future seem hard to sell to taxpayers and large emitters today. However, there are large and immediate economic benefits from reducing CO_2 emissions, such as a reduction in damages from air pollution, energy efficiency and impact on competitiveness (Hamilton, 2017). McKinsey's greenhouse gas abatement curve is one of the more advanced approaches to estimate abatement cost with a time horizon until 2030. It shows the trade-off between cost and abatement potential. The model maps the impact of abatement measures (CO_2 saved through each measure) with their net cost, limited to measures which do not cause more than 60 Euro in net cost per ton of CO_2 eliminated (McKinsey & Company, 2013). Some of the measures (most of them related to energy efficiency or land management) show a positive economic outcome until 2030. In fact, the total economic cost of implementing all options of the model would be less

than 1% of the global GDP by 2030, totaling 47 billion tons of CO_2 saved per year (Ritchie, 2017). Additionally, most positive economic long-term effects are hard to estimate, such as the health impact of reducing air pollution in China. Also, it is likely that further technology advancements and innovations will reduce the cost of abatement (Zenghelis, 2018). However, the highest-impact measures are typically on the expensive side of the scale, combining an initially high capital intensity of investment with a decrease of operating costs spread over the following years. For all measures combined, upfront capital investments of 530 billion Euro per year by 2020 and 810 billion Euro per year by 2030 would be required (Ritchie, 2017).

Given the benefits of CO_2 reductions outweighing the cost and given the increased pressure on policy makers and carbon emitters – why is progress in CO_2 reductions not taking place in an accelerated manner? Arguably, because the magnitude of benefits many times is unclear and the benefits cannot be distributed to one party (in particular, the party where the cost occurred), but the measures in general tend to benefit all actors, also passive ones, called "free-riders" (Traeger, 2009).

1.4 Politics of Climate Change Mitigation

Entering the field of game theory might lead to a better understanding of the dynamics between the actors that are responsible for reducing CO_2 emissions and thus should support in drawing conclusions on what results these actors might achieve depending on the reward maximization scheme they are applying. Dyke (2016) argues that climate change offers a good setup for a so-called "general-sum stochastic game", with individual actors (be it countries, corporations, investors or consumers) depleting a common resource – the earth's capacity to absorb CO_2 – to their short-term individual financial benefit. There is thus a misalignment between incentives for the individual and the group. In game theory, a game in which individuals collectively deplete a common-pool resource is called the "tragedy of the commons". The existence of this sort of game in climate change is currently fostered by the absence of a multinational regulatory system that would reward cooperation and punish free riding caused by a reward maximization scheme that is purely based on short-term self-interest (Dyke, 2016).

While in former critical political situations, locally induced approaches to cooperation have shown over time to build trust between actors and enable large-scale cooperation and social norms that benefit the collective, the issue arising in climate change is the difference in magnitude and timeframe of the negative impact of climate change that individual nations and groups of nations face. While a developing country in the midst of the sea rather tends to face a strong negative impact in a mid-term scenario of further rising sea levels (one such example is Papua New Guinea), a landlocked and developed country with a moderate climate and a strongly service and industrial sector-oriented economic output might not face significant negative impact by climate change in the near future (Kelman, 2015). Additionally, the economic operations of these developed and industrialised countries tend to have the largest impact on the climate, with these countries remaining least affected in the mid-term (Dyke, 2016). In game theory, when all parties are to choose an action considering the probability of the actions that other parties take given their assumed reward maximization scheme, they will choose the action that maximizes their individual benefit, given the actions that they assume the other parties to take. When, in a non-cooperative game, neither party can change their action without foregoing some of their individual benefit, we call this a "Nash equilibrium". In the case of climate negotiations, collectively postponing action has been the norm Nash equilibrium in the past and present, even though not being efficient (meaning not maximizing the cumulative benefit for all parties, Traeger, 2009).

Previous policy attempts could not proceed past the soft law stage, leaving it open for the parties to not participate at all or to later decide to not comply with the rules (Traeger, 2009). As the effects of

climate change move closer, non-Nash strategies, like local alliances and cooperation, need to enter main-stage in order to prevent further future damage (Mond, 2013; Traeger, 2009), given that global cooperation or law enforcement strategies have failed so far. Progressive policy making could also foster to break out of the Nash equilibrium, by making individual actions rational, for instance through a global fund that benefits those countries, corporations and subsequently individuals, that actively reduce CO_2 emissions (Kenyon, 2018). Another approach is linking the commitment to international environmental soft law with the membership in important multinational organizations (like WTO or EU), such that free-riding parties are potentially excluded from these organizations, as proposed by Traeger (2009). Kelman (2015) criticizes that climate change is tackled separately from the other sustainable development goals (SDG), and that a more integrated and interdisciplinary policy approach needs to be implemented.

2 Multi-Agent Reinforcement Learning and Mixed Markov Games

In the following chapter, we will enter the world of reinforcement learning and explore how multi-agent reinforcement learning might help us to draw important insights to solve the problem of climate change.

2.1 Reinforcement Learning & Q-Learning

In reinforcement learning, an agent learns to behave in order to maximize its cumulative reward in an environment by choosing some action from a (finite or infinite) action space and given an initial state, receiving a reward and observing a subsequent transition of the state. Such a process is called a "Markov Decision Process" (MDP). The goal of an agent in an MDP is to maximize its discounted cumulative reward (defined by a reward function). By actively pursuing its goal, the agent either directly follows or subsequently creates a policy, which maps potential states to the actions that the agent is assumed to take to maximize its reward. While a deterministic policy maps a state to a single best action, a stationary policy assigns a probability distribution over actions to be taken to each state (Littman, 2000). In an environment that does not arrive with a pre-set policy, so-called "off-policy" methods can estimate the value of each action given a particular state, the so-called "action-value" (Busoniu et al., 2010).

Q-Learning is a model-free, off-policy action-value estimation method that always converges to the optimal policy, based on a so-called "Q-table", consisting of all possible states (rows) and actions (columns) in an environment, showing the result of the so-called "Q-function" for each given state-action pair. The Q-function, denoted as

$$Q^*(s,a) = r(s,a) + \gamma * Q * (s',a)$$

, calculates Q-values, representing the reward r of an action a in a given state s, plus the maximum reward achievable by the best action a' in the next state s', discounted by the factor gamma γ (temporal-difference learning). Q* denotes the highest Q-value for a given row, thus defines the action in a given state that achieves the highest Q-value (Valkov, 2017).

The Q-table is updated after each iteration by

$$Q(s,a) = (1-\alpha) * Q(s,a) + \alpha * (r + \gamma * Q^*(s',a))$$

, where α denotes the learning rate of the algorithm and $Q^*(s',a)$ denotes the value of the next state as the maximal Q-value for the best action in the next state (Kansal & Martin, 2018).

In order for the agent to not just exploit the best action based on previous observations and the learned Q-values, but also keep exploring the other options in order to receive Q-values for the whole action space, agents are adjusted to sometimes take randomized actions, as controlled by the parameter epsilon ε (Littman, 2000).

2.2 Multi-Agent Reinforcement Learning

When multiple agents apply reinforcement learning in a shared environment, their optimal policies are additionally dependent from the other agents' actions and policies (Nowé, 2014).
Multi-agent Q-Learning (MultiQ) is the concept of extending the Q-function by not just mapping all possible states with all possible actions of one agent, but with the potential combination of actions of all agents in the game. Multi-agent Q-Learning thus faces the challenge of the "curse of dimensionality", meaning an exponential increase of required storage and computing power for each additional modelled state and agent (Greenwald & Hall, 2003). An alternative for modelling multiple agents in a Markov Decision Process and approximating their optimal Q-values is to rely on single-agent Q-learning. In this scenario, the Q-functions for each agent are calculated separately, based on their own actions only. This means that the actions of the other agents do not directly influence the Q-values of a given agent. However, depending on the model, the shared state of the agents can influence the actions of the individual agents, thus leading to an indirect dependence between their actions, as all their actions impact the shared state (Busoniu et al., 2010).

2.3 Mixed Markov Games & Correlated Equilibria

A game containing agents that are neither fully collaborative nor fully competitive is called a "Mixed Game". A mixed game within a dynamic environment, where the agents optimize an MDP as independent learners and in a decentralized manner, is called a "Mixed Markov Game" (MG). As such, the MG is a general-sum game, meaning that receiving rewards is not mutually exclusive and thus the benefit of one agent is not necessarily to the disadvantage of the other agents (which would be a zero-sum game, Pérolat, 2016).

While the pursuit of a Nash equilibrium is not always efficient (Traeger, 2009), the concept of a "correlated equilibrium" generalizes the Nash equilibrium and allows the agents to both make their decision dependent on the other agents' decisions, but also to optimize according to their own reward function. Correlated-Q Learning applies the correlated equilibrium to Markov Games. While agents are still allowed to optimize their reward functions individually (based on single-agent Q-Learning), their reward function itself is subject to consideration of the other agents' rewards in an equilibrium policy (Greenwald & Hall, 2003). Greenwald and Hall define four different equilibrium policies that each maximize the sum, maximum or minimum of either each of the individual agents', some of the individual agents' or the cumulative reward. An equilibrium policy that maximizes the sum of the agents' rewards is called a "utilitarian policy" (Greenwald & Hall, 2003).

3 Approach

This study models the behavior of multiple agents representing groups of nation states in reducing their CO_2 consumption while taking actions selfishly but not competitively, in a mixed Markov game setting with a global state and correlated reward functions and policies, modelled based on Coordination Equilibrium Q-Learning (CE-Q) reaching a utilitarian equilibrium, and reinforcement

learning. The goal of the Markov game is to overcome the "Tragedy of the commons", where multiple agents deplete a natural resource by purely maximizing their individual benefit (Dyke, 2016). The simulation is written in Python programming language, using only "NumPy" and "Random" as additional Python packages. The simulation is executed in Google Colab, which allows any spectator to run the simulation in their browser without having to install any dependencies.

The following simulation setup is chosen:

- The agents play a mixed Markov stochastic game with a fully observable environment and a known state transition function.
- There are three agents, defined as groups of countries. The first agent is not estimated to be immediately affected by the effects of global warming, but only in the longer term. Let us call it the "long-term impact agent" (LT). The second agent experiences a slight impact in the short-term and is equally affected in the longer term. Let us call it the "mid-term impact agent" (MT). The third agent is estimated to be already affected by the effects of global warming in the shorter term and equally affected in the longer term. Let us call it the "short-term impact agent" (ST).
- There is one global state which is initialized and defined as a small normalized scalar number representing the real global CO_2 consumption of t_0 (the starting year of the game). These are 4.97 tons CO_2 per capita as of 2014, as defined by the World Bank (2018).
- The number of state transitions (between two states or "periods"), representing the number of years the game is played, is chosen by the spectator of the simulation. The spectator can choose between 1 and 30 years.
- The number of learning epochs of the algorithm is also defined by the spectator of the simulation. The spectator can choose between 1 and 50 learning epochs.
- The agents undertake parallel actions which are defined as the per capita reduction in the CO_2 consumption of the agents. The action space is defined as (-0.2, -0.16, -0.12, -0.08, -0.04, 0, 0.04, 0.08, 0.12, 0.16, 0.2). LT has a higher (but non-greedy) tendency for exploitative actions, whereas ST has a higher tendency for explorative actions, as defined by their respective epsilon ε values. The actions are chosen from a finite action space. The epsilon values are decayed over time. This allows the agents to explore larger parts of the action space (Busoniu et al., 2010).
- The state transition function for a given period is defined as the global state before that period, minus the average per capita CO_2 consumption reduction achieved by all agents in that period.
- The immediate rewards of the agents are defined based on the per capita CO_2 consumption reduction achieved by the agent, multiplied by a reward factor (scalar number between 0 and 1), which varies depending on the agent, and subtracted by the cost of action, which is defined as the per capita CO_2 consumption reduction and is multiplied by a cost factor, which is defined as 5.
    - LT receives a lower scalar reward factor, thus a smaller immediate reward. This is because the long-term agent does not necessarily feel the negative impact of climate change directly. Its motivation to invest in CO_2 emission reductions – the agent considering mid- to long-term impact effects of investment decisions – is lower.
    - On the other end of the spectrum, ST receives a higher scalar reward factor, thus a higher immediate reward. This is because the short-term agent does feel the negative impact of climate change sooner. Its motivation to invest in CO_2 emission reductions is higher.

- The discounted future reward of the next state is modelled based on Q-learning. The discount factor of the future reward γ is the same for all agents, given that all agents face the same long-term effects. The learning rate α is decayed over time (Bowling & Veloso, 2001).
- The cumulative reward is defined as the sum of the immediate rewards of all agents in the respective period, plus the sum of the per capita CO_2 consumption reductions achieved by the agents in the respective period, multiplied by a scalar factor.
- Three different policies emerge from training the agents in the simulation – selfish, greedy and utilitarian policies. These policies are compared regarding the cumulative reward they achieve, as well as the per capita CO_2 consumption reduction they achieve, expressed by the final global state in the end period. In the selfish policy, the agents choose those actions which they have learned to maximize the sum of their respective individual immediate and future rewards. Under greedy policies, the agents choose those actions that maximize their respective individual immediate rewards in the respective period. Finally, under utilitarian policies, the agents choose those actions which maximize the cumulative reward over all periods.

The simulation and source code are open to the public and available under the links specified in the "Appendix" section. The simulation can be run in the browser without the need to install any dependencies. The results can thereby be reproduced and observed by each reader willing to do so.

4 Results & Discussion

The simulation is tested with six different total period lengths (5, 10, 15, 20, 25 and 30 periods), in 50 learning epochs and five test runs each. The following table sums up the final cumulative rewards and global states for the three different policies, "Utilitarian", "Selfish" and "Greedy", averaged over the five test runs per period configuration. The cumulative reward is defined by the cumulative reward function, the final global state is defined as the final average per capita CO_2 consumption of all agents combined.

		5 Periods	10 Periods	15 Periods	20 Periods	25 Periods	30 Periods
Cumulative reward	Utilitarian	*11.04*	*24.21*	*34.19*	*48.46*	*64.37*	*73.31*
	Selfish	*7.44*	*16.17*	*16.87*	*28.96*	*25.64*	*33.78*
	Greedy	*10.19*	*17.28*	*23.02*	*28.07*	*39.43*	*44.87*
Final global state	Utilitarian	*4.26*	*3.60*	*2.70*	*1.93*	*0.86*	*0.18*
	Selfish	*4.44*	*3.97*	*2.95*	*2.77*	*1.40*	*0.92*
	Greedy	*4.62*	*4.05*	*3.35*	*2.61*	*2.37*	*1.67*

When looking at above-shown table, as well as the policies and the Q-tables of the three agents (samples can be found in the "Appendix" section), the following observations can be inferred.

Cumulative reward

- The cumulative reward is highest in each period configuration when utilitarian policies are applied by the agents.
- The difference in cumulative rewards between utilitarian policies and other policies grows with an increasing number of periods simulated.
- Greedy policies consistently achieve higher cumulative rewards (with one exception) than selfish policies.
- Selfish policies show an abnormal decrease in cumulative reward achieved when modelled for 20 and 25 periods, compared to the other period configurations.

Final global state

- Utilitarian policies consistently achieve lower global states than the other policies.
- Selfish policies consistently achieve lower global states (with one exception) than greedy policies.

Policies and actions

- Utilitarian policies tend to imply more closer-to-maximum ("-0.2") actions than the other policies, while selfish and greedy policies show similar tendencies.
- All three types of policies of all agents show to converge towards the maximum action "-0.2" with an increasing number of periods applied.
- Agent ST tends to consistently take more action than agent MT, which tends to take more action than agent LT.

Q-tables

- For 20 and more periods modelled, the Q-tables of all agents show an increasing sparsity which is due to the increasing number of state(period-)-action pairs to be explored.

5 Conclusions & Future work

The results show that utilitarian policies achieve both higher cumulative rewards and a lower per capita CO_2 consumption, than selfish and greedy policies. In order to realize utilitarian policies, the current soft law approach on international climate policy needs to be enhanced by local commitments of smaller groups of countries and increased political and economic pressure. The policies of the modelled agents show that significantly reducing the global per capita CO_2 consumption is not an option, but a rational thing to do, both from an economic and ecological perspective. Bringing these two worlds together, the cost of carbon emissions needs to be internalized in our economic systems, to incentivize both direct polluters and consumers to rethink their CO_2 choices. While developed, landlocked countries might not face the impact of climate change as much in the near term than a developing island like Papua New Guinea, they will have to face equal consequences in the further future. Multi-agent reinforcement learning can help us to model and compare the economic and political dynamics and implications of new approaches to climate policy.

Future work could build on more sophisticated and thus realistic models, including the use of big data for function approximation and more sophisticated, sensitive and dynamic functions. Some of the potential areas of improvement include but are not limited to: the number and diversity of agents, the cost and variety of mitigation measures, the different cost and impact of these measures depending on which period they are applied to, the state transition function, mapping the global state (the global CO_2 consumption) to its impact (the consequences of global warming and climate change), mapping this impact to the agent's reward functions, the reward functions in general, modelling the environment in a partially observable MDP which includes the additional mapping from observations to states, more sophisticated models to change the parameters depending on time, the use of different MARL algorithms depending on the model (such as Team-Q, WoLF-PHC, SG-SP, different variations of CE-Q or Nash-Q, to name just a few), accounting for models of cooperation between agents and finally, modelling real-time parallelization of and communication between agents (e.g. through using osBrain in the case of Python).

This study has intended to show how Multi-Agent Reinforcement Learning (MARL) and game theory can be applied to explain the current issues in climate change negotiations and CO_2 consumption

reduction and how this currently and increasingly popular AI technology can contribute to solving these issues in the future. The study has also intended to symbolize that reinforcement learning can be applied beyond its undoubtedly important application area of physical games and simulations and that once combined with game theory, it has a great potential for achieving groundbreaking insights about human cooperation and competition, and thus for economic and political decision making.

6 Acknowledgements

The author thanks the School of AI and its Director, Siraj Raval, for their support, as well as Kushal Sharma for his help and contributions.

References

Bowling, M., & Veloso, M. (2001, August). Rational and convergent learning in stochastic games. In International joint conference on artificial intelligence (Vol. 17, No. 1, pp. 1021-1026). Lawrence Erlbaum Associates Ltd.

Budd, C. (2016). Climate change: Does it all add up? https://plus.maths.org/content/climate-change-does-it-all-add

Busoniu, L., Babuska, R., & De Schutter, B. (2010). Multi-agent reinforcement learning: ˇAn overview

Carbon Disclosure Project (CDP, 2017). The Carbon Majors Database - Carbon Majors Report 2017. https://b8f65cb373b1b7b15feb-c70d8ead6ced550b4d987d7c03fcdd1d.ssl.cf3.rackcdn.com/cms/reports/documents/000/002/327/original/Carbon-Majors-Report-2017.pdf?1499691240

Dyke, J. (2016). Can game theory help solve the problem of climate change? https://www.theguardian.com/science/blog/2016/apr/13/can-game-theory-help-solve-the-problem-of-climate-change. The Guardian

Greenwald, A., Hall, K., & Serrano, R. (2003, August). Correlated Q-learning. In ICML (Vol. 3, pp. 242-249).

Hamilton, K. (2017). Economic co-benefits of reducing CO2 emissions outweigh the cost of mitigation for most big emitters. http://www.lse.ac.uk/GranthamInstitute/news/economic-co-benefits-of-reducing-co2-emissions-outweigh-the-cost-of-mitigation-for-most-big-emitters/. The London School of Economics and Political Science – Grantham Research Institute on Climate Change and the Environment

Intergovernmental Panel on Climate Change (IPCC, 2018). Less Than 2 °C Warming by 2100 Unlikely. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6070153/

Intergovernmental Panel on Climate Change (IPCC, 2013). Fifth Assessment Report (AR5). https://www.ipcc.ch/report/ar5/

Kansal, S., & Martin, B. (2018). Reinforcement Q-Learning from Scratch in Python with OpenAI Gym. https://www.learndatasci.com/tutorials/reinforcement-q-learning-scratch-python-openai-gym/

Kelman, I. (2015). Joint action on climate change: Facts and figures. https://www.scidev.net/global/climate-change/feature/joint-action-climate-change-facts-figures.html

Kenyon, R. (2018). How to break the climate change Nash equilibrium. https://medium.com/nori-carbon-removal/how-to-break-the-climate-change-nash-equilibrium-f4ca3354cb8b

Littman, M. L. (2001). Value-function reinforcement learning in Markov games. Cognitive Systems Research, 2(1), 55-66.
McKinsey & Company (2013). Pathways to a Low-Carbon Economy. https://www.cbd.int/financial/doc/Pathwaystoalowcarboneconomy.pdf
Mond, D. (2013). Game Theory and Climate Change. http://homepages.warwick.ac.uk/~masbm/ClimateCourse/MondTalks/climategameunpause.pdf. University of Warwick, Mathematics Institute
Nowé, A., Vrancx, P., & De Hauwere, Y. M. (2012). Game theory and multi-agent reinforcement learning. In Reinforcement Learning (pp. 441-470). Springer, Berlin, Heidelberg.
Pérolat, J., Strub, F., Piot, B., & Pietquin, O. (2016). Learning Nash Equilibrium for General-Sum Markov Games from Batch Data. arXiv preprint arXiv:1606.08718.
Ritchie, H. (2017). How much will it cost to mitigate climate change? https://ourworldindata.org/how-much-will-it-cost-to-mitigate-climate-change. Our World in Data.
Talbot, D. (2014). Carbon Sequestration: Too Little, Too Late? https://www.technologyreview.com/s/531531/carbon-sequestration-too-little-too-late/
The World Bank (2018). CO2 emissions (metric tons per capita). https://data.worldbank.org/indicator/EN.ATM.CO2E.PC?end=2014&start=1960&view=chart&year_high_desc=true
Traeger, C. (2009). The economics of climate change. International Cooperation and Climate Policy. https://are.berkeley.edu/~traeger/Lectures/ClimateChangeEconomics/Slides/7%20International%20Cooperation%20-%201%20A%20Game%20Theoretic%20Perspective.pdf. UC Berkeley
Union of Concerned Scientists UCS USA (2017). CO2 is the issue: https://www.ucsusa.org/global-warming/science-and-impacts/science/CO2-and-global-warming-faq.html#.W_aH42hKg2x
United Nations Framework Convention on Climate Change (UNFCCC) (2018). https://unfccc.int/process/the-paris-agreement/what-is-the-paris-agreement
United Nations (2011). The Paris Agreement. World Economic and Social Survey 2011. http://www.un.org/en/development/desa/policy/wess/wess_current/2011wess_chapter2.pdf
Valkov, V. (2017). Solving an MDP with Q-Learning from scratch. https://medium.com/@curiousily/solving-an-mdp-with-q-learning-from-scratch-deep-reinforcement-learning-for-hackers-part-1-45d1d360c120
Zenghelis, D. (2018). How much will it cost to cut global greenhouse gas emissions? http://www.lse.ac.uk/GranthamInstitute/faqs/how-much-will-it-cost-to-cut-global-greenhouse-gas-emissions/. The London School of Economics and Political Science – Grantham Research Institute on Climate Change and the Environment

A Appendix

A.1 Links to Source Code & Dynamic Paper

Link to Source Code:
https://colab.research.google.com/drive/1wgy766JmPyX8wQhr2pyZpKRfyEpvH6no

Link to Dynamic Paper:
https://colab.research.google.com/drive/1wOS8w3V6DoSO-c46Pbq-OUr7kQl8cFl9

A.2 Sample Simulation Results

The following figure shows sample utilitarian, greedy and selfish policies of agent MT after 50 learning epochs and 10 periods, representing the respective action that achieves the highest Q-value for each period.

```
MT's Strategy to achieve Highest Cumulative Reward:
 [-0.2, -0.08, -0.2, -0.2, -0.2, -0.2, -0.16, -0.16, -0.2, -0.2]
MT's Strategy to achieve Highest Immediate Reward:
 [0.04, 0.08, -0.2, -0.04, -0.2, -0.2, 0.04, 0.12, -0.2, -0.2]
Selfish Policy of MT, based on MT's Final Q-Table:
 [0.08, 0.04, 0.08, -0.2, -0.04, 0.08, -0.2, 0.04, 0.12, -0.04, -0.2]
```

The following figure shows a sample learned Q-table of agent MT after 50 learning epochs and 10 periods, showing the respective period in the left column, the respective action from the action space in the top row and the respective Q-values for each learned period/action combination ("0." if not learned after 50 epochs).

```
Final Q-Table of MT:
       -0.2 -0.16 -0.12 -0.08 -0.04  0    0.04  0.08  0.12  0.16  0.2
p1 [ 0.    0.06 -0.02  0.17 -0.12  0.16 -0.05  0.19 -0.   -0.09  0.03]
p2 [-0.17 -0.08  0.05 -0.16  0.   -0.07  0.2   0.02 -0.05  0.1  -0.04]
p3 [-0.   -0.02  0.07  0.07 -0.11  0.05  0.01  0.1   0.   -0.04  0.06]
p4 [ 0.39  0.04 -0.03 -0.08 -0.02  0.    0.    0.06  0.03  0.15  0.04]
p5 [-0.08  0.02 -0.09  0.1   0.16  0.02 -0.02  0.   -0.07 -0.04 -0.04]
p6 [ 0.08 -0.06 -0.05  0.    0.15  0.14  0.03  0.17 -0.08 -0.01  0.  ]
p7 [ 0.3   0.04 -0.03 -0.05  0.02  0.1   0.05  0.19  0.    0.07 -0.04]
p8 [ 0.03  0.    0.    0.11  0.07 -0.09  0.12  0.   -0.01 -0.06  0.  ]
p9 [-0.13  0.    0.    0.   -0.07  0.    0.   -0.11  0.07 -0.    0.04]
p10 [ 0.02  0.02 -0.07  0.01  0.05 -0.01 -0.03 -0.05 -0.02 -0.09  0.02]
```