Learning against learning : evolutionary dynamics of reinforcement learning algorithms in strategic interactions

Summary

Computer programs automate increasingly complex tasks. Previously, tasks could be predefined, e.g., for industrial robotics. In contrast, new application domains like automated stock trading require highly adaptive agents that learn in dynamic environments and against adversarial opponents. While automated trading is increasingly adopted and now generates about a third of all trading volume in the UK, the understanding of systems in which agents are *learning against learning* is limited. The lack of a formal framework makes assessing the stability of these crucial systems practically impossible. This dissertation addresses the need for a formal framework to analyze multi-agent learning, drawing on an established relationship between multi-agent reinforcement learning and evolutionary game theory.

Previous work has shown that the behavior of stochastic multi-agent learning algorithms with an infinitesimal learning rate can be described by deterministic dynamical systems. This approach makes it possible to employ tools from dynamical systems theory to judge the convergence properties of learning algorithms in strategic interactions. In particular, the dynamics of Q-learning have been related to an extension of the replicator dynamics from evolutionary game theory with an additional exploration term. However, this equivalence is based on the simplifying assumption that all actions are updated at every time step. Here, I show that this leads to a discrepancy between the observed algorithm performance and the idealized evolutionary model. Since the idealized model shows preferable behavior, I introduce the variation *Frequency Adjusted Q-learning* (FAQ-learning) that adheres to the idealized dynamics. In addition, this solidified link is used to provide a convergence proof for FAQ-learning in two-agent two-action games. In the limit of infinite time, FAQ-learning converges to stable points whose distance to Nash equilibria is related to the degree of exploration of the algorithms. Hence, this proof relates multi-agent reinforcement learning to evolutionary and classical game theory.

In subsequent chapters, I extend the evolutionary framework for multi-agent learning to more realistic settings, like multiple states and varying exploration rates. Furthermore, I introduce an orthogonal visualization of the dynamical systems that provides a method to design time-dependent parameters of agents (e.g., exploration) and games. The evolutionary game theoretic models have the replicator dynamics as a common building block, and a similar term appears in the dynamical systems describing *Infinitesimal Gradient Ascent* (IGA). The commonalities and differences between variations of IGA dynamics and replicator dynamics are discussed in detail. In essence, the difference depends on whether the payoff signal is known for all actions at every time

step or whether it needs to be sampled for one action at a time. This implies that the reinforcement-learning algorithms can be seen as stochastic gradient ascent on the payoff function. The comparative discussion of these two independently developed approaches unites them under the same terminology and provides a basis for further cross-fertilization.

Finally, the merits of an evolutionary analysis are demonstrated in two application domains: auctions and poker. The analysis critically evaluates strategic behavior and compares the results with domain knowledge. The strategic payoffs from the application domains are captured in a *heuristic payoff table* by observing various finite strategy constellations. Subsequently, the expected payoff for an arbitrary mix of strategies in an infinite population can be approximated from the heuristic payoff table, and is used in the context of the evolutionary dynamics. In poker, results are in line with expert advice, even more so if exploration is accounted for in the evolutionary model. Similarly, results in simulated double auctions confirm results from previous work. More specifically, performance in double auctions does not increase monotonically with more information about the future price development: traders with no information perform at market average, while traders with little information are exploited by insiders with a lot of information; this results in a J-curve for the value of information. If information comes for free, insiders drive other traders extinct. If on the other hand information is costly, less informed traders may prevail. This work provides a good basis to study the resilience to exogenous events, like trader in- and outflow, that may further disturb the system.

Overall, this dissertation contributes to the state-of-the-art in multi-agent reinforcement learning in several ways: (1) a critical evaluation and improvement of the link between Q-learning and its idealized dynamics enables a proof of convergence for the variant Frequency Adjusted Q-learning, (2) the evolutionary framework is extended to more realistic settings and enriched by new perspectives, and (3) application domains demonstrate how practical insights can be derived from the theoretical models. Tying together tools from reinforcement learning, dynamical systems, evolutionary and classical game theory, this dissertation lays out a formal framework for the analysis of systems in which agents are learning against learning, paving the way for many viable future research endeavors.

Samenvatting

In dit proefschrift bestudeer ik computerprogramma's (agenten) die samen leren te coördineren of te concurreren. Er wordt hoofdzakelijk onderzocht hoe hun leerprocessen elkaar beïnvloeden. Dergelijke adaptieve agenten spelen reeds een belangrijke rol in onze maatschappij. Zo nemen geautomatiseerde agenten bijvoorbeeld al deel aan de financiële handel en genereren in een aantal Amerikaanse markten reeds meer transacties dan de mens. Ondanks de grootschalige toepassing is het voor de meerderheid van leeralgoritmen enkel bewezen dat ze goed presteren als zij geïsoleerd optreden—zodra een tweede agent invloed heeft op de omgeving of uitkomsten, zijn de meeste garanties niet meer van toepassing. Mijn belangrijkste bijdragen zijn de uitbreiding en de toepassing van methodiek om te beoordelen in hoeverre optimaal gedrag in strategische interacties door leeralgoritmen wordt benaderd. Het gedrag van deze algoritmen wordt geformaliseerd op basis van stochastische en dynamische systemen, en hun korte en lange termijn prestaties worden in het kader van de klassieke en evolutionaire speltheorie besproken.

Zusammenfassung

In dieser Dissertation werden Computerprogramme (Agenten) analysiert, die mit- und gegeneinander lernen. Im Besonderen wird darauf eingegangen, wie sich die Lernprozesse der Agenten gegenseitig beeinflussen. Solche adaptive Agenten spielen schon heute in verschiedenen Bereichen unseres Lebens eine ausschlaggebende Rolle, auch wenn dies häufig übersehen wird; so nehmen z.B. Computer-Agenten an Finanzmärkten teil und generieren in einigen US Märkten größere Transaktionsvolumen als menschliche Händler. Für allein agierende lernende Agenten bzw. deren Lernverfahren können Konvergenz zum optimalen Verhalten und dessen Stabilität häufig garantiert werden. Solche Garantien sind für Systeme, bestehend aus mehreren, interagierenden, lernenden Agenten, im Allgemeinen nicht übertragbar, da das optimale Verhalten (das Lernziel) des einen Agenten vom Verhalten der anderen Agenten abhängt und sich fortwährend ändern kann. In der vorliegenden Dissertation wird eine Methode entwickelt und angewandt, die es erlaubt zu bewerten, inwiefern sich interagierende Lernverfahren an das theoretisch erreichbare Optimalverhalten in strategischen Konflikten annähern. Das Verhalten dieser Lernverfahren wird mit Hilfe von stochastischen und dynamischen Systemen formal modelliert, und das Kurz- und Langzeitverhalten wird im Kontext von klassischen und evolutionären spieltheoretischen Lösungansätzen diskutiert.