**TESI DI LAUREA in Ingegneria Informatica, percorso ICT Engineering, Business and Innovation**

**MULTI AGENT LEARNING WITH LEGO® SPIKE PRIME EDUCATION**

**Index**

1. Abstract (Italian and English)
2. Introduction
3. Multi Agent Reinforcement Learning (State of the Art)
   1. Reinforcement Learning, what it is.
   2. Multi Agent Reinforcement Learning
   3. Multi Agent RL Applications
   4. Multi Agent RL Limits
4. Formal Model Definition
   1. Definition of the Markov Game for multi agent learning.
   2. Definition of an MDP.
   3. Demonstration that in centralized settings, two MG converge in a MDP.
5. The Framework Used
   1. LEGO® Spike Prime Education.
   2. Explanation of the hardware and its possibilities.
   3. Beyond Modular Programming and Pybricksdev Framework.
   4. Main action used on the robot.
6. Research and Experimentation
   1. Grid-World definition
   2. Multi-Agent Q-Learning, results and observations.
   3. Centralized Q-Learning, results and observations.
   4. Multi-Agent FAQ-Learning, results and observations.
7. Conclusion

**Multi Agent Reinforcement Learning (State of the Art)**

**Reinforcement Learning**

Al fine di comprendere al meglio gli argomenti trattati all’interno di questa tesi, è doveroso presentare il punto cardine intorno al quale ruota la sperimentazione e le successive conclusioni che verranno presentate all’interno di questo documento, l’apprendimento basato su Reinforcement Learning.

Allo Stato dell’Arte, il Reinforcement Learning è un sottotipo di Unsupervised Learning nell’ambito dell’apprendimento automatico in cui un’agente impara a compiere azioni all’interno di un ambiente ignoto al fine di massimizzare una funzione di ricompensa. Esso appartiene al ramo del Machine Learning chiamato Unsupervised Learning proprio perché gli input del modello sono ignoti e vengono appresi con “l’esperienza’ stessa dell’agente all’interno dell’ambiente.

Al fine di definire al meglio il Reinforcement Learning, è necessario introdurre diversi elementi che lo compongono:

* **Agente**: l'entità che apprende e prende decisioni. Può essere un programma, un robot o qualsiasi sistema in grado di interagire con un ambiente.
* **Ambiente**: il contesto in cui l'agente opera. Può essere completamente o parzialmente osservabile dall’agente.
* **Azioni**: il set di mosse che un agente può prendere una volta osservato lo stato corrente dell’ambiente.
* **Stato**: la rappresentazione dell'ambiente in un dato momento. Lo stato è ciò che l'agente osserva per prendere decisioni.
* **Feedback o Reward**: un valore numerico che l'agente riceve dopo aver compiuto un'azione in uno stato specifico. L'obiettivo dell'agente è massimizzare la somma cumulativa delle ricompense nel tempo.

Il processo di apprendimento nel Reinforcement Learning coinvolge l'agente che esplora l'ambiente, apprende quali azioni sono più vantaggiose in determinati stati e aggiorna la propria strategia di decisione per massimizzare la ricompensa accumulata.

**Multi Agent Reinforcement Learning (MARL)**

Ma, come si evince dal titolo stesso di questo elaborato, il nostro focus verterà su una branca specifica del Reinforcement Learning, ovvero il Multi Agent Reinforcement Learning.

Per capire al meglio le impostazioni di questo tipo di problema d’apprendimento, è necessario introdurre i sistemi multi-agente.

Un sistema multi-agente è definito come un gruppo di entità autonome che interagiscono tra loro e condividono lo stesso ambiente, che usualmente percepiscono attraverso sensori e in cui si muovono attraverso attuatori, ma non solo.

Sebbene un agente all’interno di questo tipo di sistemi possa essere programmato per comportarsi in un determinato modo in precedenza, risulta comunque necessario far si che lui apprenda parte del suo comportamento “online”, in modo tale da migliorare le singole prestazioni, ma anche quelle di tutto l’ambiente. Ciò avviene solitamente perché la complessità dell’ambiente rende difficile, o addirittura impossibile, la progettazione a priori di una policy ottima. *( L. Busoniu, R. Babuska and B. De Schutter, "A Comprehensive Survey of Multiagent Reinforcement Learning," in IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), vol. 38, no. 2, pp. 156-172, March 2008, doi: 10.1109/TSMCC.2007.913919*)

Al fine di garantire questo apprendimento “online”, è necessario appunto ricorrere a specifici algoritmi che si rifanno ai concetti di esplorazione e valorizzazione del sopra citato Reinforcement Learning, introducendo quindi il Multi-Agent Reinforcement Learning (MARL).

La presenza di agenti multipli modifica di gran lunga le possibilità di apprendimento poiché ogni singolo agente è costretto ad “imparare” considerando anche altri agenti che si muovono nel suo stesso ambiente.

In un ambiente MARL gli agenti interagiscono tra di loro in diversi modi: essi devono imparare a prendere decisioni in modo cooperativo o competitivo a seconda dell’obiettivo prefissato e massimizzare, appunto, le loro ricompense complessive.

È per questo che, in aggiunta ai sopra citati elementi che compongono un algoritmo di Reinforcement Learning, vanno menzionati dei nuovi che risultano necessari in questo nuovo scenario:

* **Agenti Multipli**: ci sono più agenti che coesistono e interagiscono nell'ambiente. Ogni agente può avere obiettivi individuali, e l'apprendimento avviene considerando sia le azioni dell'agente stesso che quelle degli altri agenti.
* **Comunicazione e Coordinamento**: gli agenti possono comunicare tra di loro per scambiare informazioni, coordinarsi o prendere decisioni condivise. Questo aspetto è cruciale in scenari in cui il successo di un agente dipende dalle azioni degli altri.
* **Competizione o Collaborazione**: A seconda del contesto, gli agenti possono essere in competizione tra di loro per raggiungere obiettivi individuali o possono collaborare per massimizzare un obiettivo comune.
* **Problemi di Non-Stationarity**: L'ambiente è spesso non stazionario a causa delle azioni in continua evoluzione degli altri agenti, il che rende la pianificazione e l'apprendimento più complessi rispetto al caso in cui un singolo agente agisce da solo.

Questi elementi tracciano anche quelle che sono le sostanziali differenze tra un Single Agent RL e un Multi Agent RL. Difatti, se il primo può essere formalmente ricondotto ad un Processo Decisionale Markoviano (MDP) dove appunto le decisioni dell’agente in ogni singolo stato sono dettate da singole probabilità dipendenti solo dallo stato osservato; il secondo perde questa proprietà decisionale essendo influenzato anche dai comportamenti degli altri agenti, evidenziando, appunto, la non stazionarietà dei singoli stati dell’ambiente e diventando formalmente non più un MDP, bensì un Gioco Stocastico o Gioco Markoviano che trova le sua fondamenta nella teoria dei giochi. (*L. Busoniu, R. Babuska and B. De Schutter, "A Comprehensive Survey of Multiagent Reinforcement Learning," in IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), vol. 38, no. 2, pp. 156-172, March 2008, doi: 10.1109/TSMCC.2007.913919*)

Analizzeremo meglio questi formalismi nel capitolo successivo ad essi dedicato.

**Examples and Applications**

I sistemi multi-agente stanno trovando applicazioni in un’ampia varietà di settori, tra cui robotica, controllo distribuito, gestione delle risorse, sistemi di supporto decisionale collaborativo, data mining, ecc.

In questa sezione ne prenderemo alcuni come esempio proprio per comprendere al meglio quelle che sono le possibilità che algoritmi MARL possono offrire e in quali diversi ambiti possono essere applicati. Non useremo questo approfondimento per scendere nel dettaglio delle diverse implementazioni, bensì per avere un quadro generale su cui basare le nostre successive analisi.

**Google® DeepMind**

**Google DeepMind** è un primo esempio di azienda, acquistata da Google nel 2015, che ha fatto del machine learning e dell’intelligenza artificiale i suoi punti cardine. Il loro obiettivo dichiarato è proprio quello di “ Risolvere l’intelligenza e quindi utilizzarla per affrontare importanti sfide della società". ([*https://deepmind.google/*](https://deepmind.google/)).

I loro progetti legati all’intelligenza artificiale sono svariati e tutti utilizzano diversi tipi di algoritmi.

Un esempio di utilizzo di MARL è “AlphaStar”, un software in grado di giocare attivamente al videogioco StarCraft II che ha conseguito nell’Agosto del 2019, a 7 mesi dal suo lancio, il titolo di Grandmaster all’interno del videogame. All’interno del software, come scritto nella documentazione presente sul sito, sono presenti tecniche di apprendimento automatico generiche, tra cui reti neurali, self-play via Reinforcement Learning, Multi-Agent e Imitation Learning, per apprendere direttamente dai dati di gioco con tecniche general pourpose. *(“AlphaStar: Grandmaster level in StarCraft II using multi-agent reinforcement learning”, The Alphastar Team, 30 October 2019*). Le tecniche Multi-Agent risultano quindi utili al software durante l’apprendimento di policy ottime al fine di superare e sconfiggere le policy di altri agenti presenti nel videogioco. AlphaStar è un primo esempio di Adversarial Multi-Agent Learning, accompagnato però da ulteriori metodi d’apprendimento per irrobustirne l’efficacia. Insieme ad AlphaStar, altri progetti di DeepMind utilizzano MARL per i loro scopi, come AlphaGo (gioco del Go) e AlphaZero (gioco degli Scacchi).

**Drones and Field Coverage**

In March 2018, some researchers from the Cornell University proposed a cooperative and distributed Multi Agent Reinforcement Learning of drones for field coverage.

The target where Unmanned Aerial Vehicles (UAV) that needs to explore the area in the most efficient way. The proposed MARL algorithm allows UAVs to learn cooperatively to provide a full coverage of an unknown field of interest while minimizing the overlapping sections among their field of views. Here the problem was solved using a Multi-Agent Q-Learning approach together with a game-theoretical correlated equilibrium mechanism and a function approximation used to face the joint-action selection problem and the high dimensions of the state-actions field.

(Huy Xuan Pham, Hung Manh La, David Feil-Seifer, Aria Nefian, “Cooperative and Distributed Reinforcement Learning of Drones for Field Coverage”, 20 March 2018)

**Fleet Management**

In 2018, a multi-agent reinforcement learning framework was proposed to manage fleet of online ride-sharing platforms such as Uber and Lyft. The authors introduced two algorithms to allocate cars (represented as agents) to specific zones in a city divided into hexagonal grids. Computational efficiency was enhanced using geographical and collaborative contexts, reducing action spaces and avoiding conflicting movements among agents.

They promoted this solution as a way to make efficient the fleet management strategy, improving the utilization of transportation resources and also increasing the revenue and the customer satisfaction. (K Lin, R Zhao, Z Xu, J Zhou, "Efficient large-scale fleet management via multi-agent deep reinforcement learning“, 2018)

**Lithium-Ion Battery Scheduling**

Sui and Song in 2020 proposed a reinforcement learning framework for solving battery scheduling problems in order to extend the lifetime of batteries used in electrical vehicles (EVs), cellular phones, and embedded systems. The Multi-Agent perspective was obtained by considering each battery as a single agent, generating a cooperative MARL problem that optimizes the charge and the battery’s thermal profile.

This because battery pack lifetime has often been the limiting factor in many of today’s smart systems and smart charge-discharge scheduling of battery packs is essential to obtain super linear gain of overall system lifetime. (Y Sui, S Song , “A multi-agent reinforcement learning framework for lithium-ion battery scheduling problems”)

Queste erano solo alcune delle svariate applicazioni di questo tipo di algoritmi. Esse dimostrano come i campi che necessitano di miglioramenti sotto il punto di vista del Multi-Agent Learning sono molto ampi e variegati. Ciononostante, questo tipo di implementazioni presentano diversi limiti che andremo prontamente ad analizzare.

**MARL Limitations**

Una volta comprese le impostazioni e il funzionamento degli algoritmi MARL, insieme ad alcune loro possibili applicazioni, è anche possibile intravedere quelle che sono evidenti limitazioni legate direttamente a quelli che sono gli elementi aggiuntivi, rispetto ad algoritmi Single Agent, che li connotano.

La peculiarità di sistemi Single Agent è quella di poter essere ricondotti a dei Processi Decisionali Markoviani, per i quali sussiste l’assunzione markoviana che ci garantisce che l’apprendimento è strettamente legato all’ultimo stato osservato, ovvero quello presente. Questo elimina completamente l’influenza degli stati passati durante il miglioramento dei parametri d’apprendimento, tipico appunto di sistemi stazionari.

Ad esempio, all’interno di algoritmi come il Q-Learning, il più diffuso nell’apprendimento single-agent, la regola di update prescinde esclusivamente dallo stato corrente e dallo stato futuro, senza fare alcun riferimento a quelli attraversati in precedenza.

Negli algoritmi Multi-Agent la **stazionarietà** cessa d’esistere con l’introduzione di più agenti che influiscono attivamente sull’apprendimento dei restanti. Ad esempio, una policy che risulta ottimale in un certo istante per un singolo agente potrebbe non esserlo nell’istante successivo a causa di un secondo agente che influenza “negativamente” l’ambiente e gli stati ottimali per il primo, rendendoli, per esempio, inutilizzabili.

Questo elimina l’assunzione Markoviana e aumenta la complessità e la gestione dell’apprendimento, introducendo nuove sfide e nuove tematiche d’interesse al fine di ottimizzare e assicurare il processo d’apprendimento.

L’aumento del numero di agenti incide anche su altre caratteristiche d’interesse di questi algoritmi, come la **scalabilità**. L’aumento del numero d’agenti all’interno dell’ambiente non solo aumenta la complessità di gestione degli stessi, ma anche la complessità dell’ambiente stesso, facendo incrementare esponenzialmente lo spazio congiunto delle azioni possibili (dovendo considerare insieme le azioni di un numero sempre più grande di agenti) portando le risorse computazionali necessarie a crescere esponenzialmente insieme al numero di agenti.

L’ **osservabilità parziale** è un altro dei limiti legati agli ambienti MARL. La maggior parte degli algoritmi esistenti hanno come assunzione proprio la completa osservabilità di uno stato, cosa che però in esempi pratici è difficile da replicare per tutti gli agenti che potrebbero essere più o meno consapevoli del loro stato e di quello degli altri. Ad esempio, nei casi reali gli agenti possono essere provvisti di diversi sensori o essere limitati sotto quel punto di vista, rendendo la completa osservabilità dello stato impossibile.

Il problema dell’osservabilità parziale può essere risolto introducendo la comunicazione tra i diversi agenti, che diventano quindi in grado di aggiornarsi a vicenda sullo stato attuale del sistema. Questo, però, introduce nuove limitazioni legate proprio alla comunicazione tra agenti, i quali devono poter comunicare fra loro in maniera costruttiva senza cadere in situazioni di comunicazione superflua o fuorviante. (Canese L, Cardarilli GC, Di Nunzio L, Fazzolari R, Giardino D, Re M, Spanò S. Multi-Agent Reinforcement Learning: A Review of Challenges and Applications. Applied Sciences. 2021; 11(11):4948.)

Inoltre, gli algoritmi MARL presentano problematiche anche nel bilanciamento tra cooperazione e competizione, che fanno direttamente capo ad una scelta adeguata di incentivi e ricompense. Difatti, comprendere quali ricompense sono ottimali e che tipo di comportamento favoriscono tra gli agenti può richiedere analisi approfondite e robuste analisi di sensibilità sull’algoritmo. Sul lato comunicativo, è importante comprendere se un approccio centralizzato, dove un master controlla e monitora tutte le azioni e gli stati degli agenti, oppure un approccio decentralizzato dove ogni agente resta consapevole soltanto del proprio stato e delle proprie azioni, risultino i migliori e più efficienti.

Se nel primo caso, un ambiente centralizzato garantisce una completa osservabilità dell’ambiente d’apprendimento, il rischio è quello di dover gestire ambienti troppo complessi che richiedano ingenti risorse computazionali, sfociando così nella famosa “Curse of Dimensionality”; nel secondo caso, invece, l’assenza completa o parziale di comunicazione tra le due parti potrebbe risultare in un esplorazione inefficacie e inconcludente, richiedendo quindi accorgimenti sul trovare sub-ottimi di esplorazione e equilibri virtuali tra i vari agenti.

È perciò evidente come non ci sia una vera e propria standardizzazione su questo tipo di algoritmi, rendendo ogni implementazione fine a sé stessa, con la necessità di doverla personalizzare a seconda del contesto e degli agenti che andremo ad utilizzare; questo problema rende difficile la serializzazione, la replica e soprattutto la comparazione di diversi algoritmi Multi-Agent.

I limiti degli algoritmi MARL sono evidenti e complessi, ma non invalicabili. Quello che proponiamo noi attraverso questo studio di tesi è proprio quello di trovare, attraverso prove ed esperimenti, il giusto compromesso che dimostra come un algoritmo Multi-Agent possa effettivamente funzionare in determinate condizioni, monitorandone e valutandone le prestazioni, paragonandola a suoi simili per trarne le giuste conclusioni. L’obiettivo è comprendere quali di queste precedenti limitazioni possono essere superate e quali no. Prima di farlo, dobbiamo definire al meglio quelli che sono i modelli su cui si è basata la nostra ricerca e perché li abbiamo usati.

**The Agents’ Framework**

Una volta definiti i modelli su cui si baserà la nostra sperimentazione, vogliamo presentare quelli che saranno i nostri soggetti di studio, ovvero i singoli agenti che prenderanno parte all’apprendimento MARL e che si muoveranno negli ambienti sopra definiti.

Al fine di avere una sperimentazione snella e semplice, non abbiamo optato per agenti complessi quali possono essere robot da laboratorio sofisticati, con diversi tipi di attuatori e gradi di libertà, i quali avrebbero aggiunto complessità superflua al nostro lavoro e soprattutto non necessaria. È per questo che la nostra scelta è ricaduta su un modello di robot che ha come primo obiettivo quello di avvicinare ragazzi delle scuole medie al mondo della programmazione e della logica: il robot LEGO® Education Spike Prime.

Il suo utilizzo rappresenta la doppia sfida di verificare l’efficienza di algoritmi MARL e farlo su hardware che non sono stati pensati per questo tipo di lavori.

Successivamente introdurremo il robot, le sue caratteristiche e il suo funzionamento, cercando di spiegare come è stato possibile utilizzarlo nella nostra ricerca.

**LEGO® Education Spike Prime**

LEGO® Education Spike Prime è un set LEGO® formato da un set base e una sua espansione che permette di ottenere il robot completo che potete vedere in figuraX. Il set complessivo è formato da un totale di 1132 pezzi che possono essere usati per creare la “Struttura Motrice Avanzata” o altri robot più semplici che, insieme ad oggetti di scena, possono completare le sfide proposte da LEGO® stessa.

Il nostro focus però sarà sulla “Struttura Motrice Avanzata”, che per comodità chiameremo Spike, e sulle sue componenti, quella che poi è stata effettivamente usata durante la sperimentazione.

Spike è un robot che presenta due ruote motrici da 8.8 cm e due ruote libere che lo aiutano nei movimenti di curvatura e virata. Oltre alla struttura in mattoncini LEGO® che forma il telaio del veicolo, Spike presenta un variegato set di sensori e motori:

* Sensore di Distanza: incorpora “occhi” programmabili a LED e misura ad ultrasuoni distanze comprese tra 1 e 200 cm, con una precisione di ±1 cm;
* Sensore di Forza: misura forze con valori massimi di 10 Newton (~1 kg). Può fungere anche da sensore di contatto poiché in grado di rilevare anche lievi pressioni;
* 2x Sensori di Colore: è in grado di distinguere ben 8 colori con una frequenza di campionamento a 1 kHz e misurare la luce riflessa sul suo “occhio” e la luce ambientale;
* 2x Motori Angolari Medi: presentano un sensore di rotazione integrato con posizionamento assoluto e precisione fino ad 1 grado;
* 2x Motori Angolari Grandi: più stabili rispetto a quelli medi e perfetti per eseguire controlli in linea retta.

Tutti gli elementi sopraindicati presentano degli adattatori a 6 pin per essere collegati ad altri dispositivi. In particolare, possono essere collegati all’elemento centrale del sistema, che rappresenta il vero e proprio hardware di Spike, l’Hub Intelligente. Esso ha un peso di 63 g (senza batteria) e delle dimensioni di L88.0 x W56.0 x H32.0 mm, ed è caratterizzato da:

* 6 porte input/output utilizzate per collegare sensori e motori all’hub;
* 4 pulsanti, 1 pulsante centrale di avvio, 2 pulsanti direzionali e un pulsante per attivare il pairing Bluetooth;
* Connettività Bluetooth Low Energy;
* Una matrice colorata personalizzabile di dimensioni 5x5 che può anche fungere da display;
* Un giroscopio a 6 assi in grado di monitorare la posizione relativa dell’hub (3 assi dedicati al giroscopio e 3 assi dedicati all’accelerometro);
* A rechargeable lithium-ion battery;
* Uno speaker;
* Un sistema operativo MicroPython programmabile montato su un hardware che presenta:
  + Un processore a 100MHz M4 320 KB RAM 1M FLASH;
  + 32 MB di memoria per programmi, suoni e altri file;