THESIS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

Towards Optimal Algorithms For Online Decision Making Under Practical Constraints

ARISTIDE C. Y. TOSSOU

Department of Computer Science and Engineering
CHALMERS UNIVERSITY OF TECHNOLOGY

Gothenburg, Sweden 2019

Towards Optimal Algorithms For Online Decision Making Under Practical Constraints
ARISTIDE C. Y. TOSSOU

ISBN 978-91-7905-207-2

Doktorsavhandlingar vid Chalmers Tekniska Högskola
Ny serie nr. 4674
ISSN 0346-718X
Technical Report No. 177D
Department of Computer Science and Engineering
Data Science and AI Division

Department of Computer Science and Engineering
Chalmers University of Technology
SE-412 96 Gothenburg
Sweden
Telephone: +46 (0)31-772 1000

Cover:
An autonomous drone wondering if it should go down the rock or fly.

Chalmers Reproservice
Gothenburg, Sweden 2019

To Rose, Armand, Jean-Luc and Bernadette.

Towards Optimal Algorithms For Online Decision Making Under Practical Constraints

Aristide C. Y. Tossou

Department of Computer Science and Engineering
Chalmers University of Technology

Abstract

Artificial Intelligence is increasingly being used in real-life applications such as driving with autonomous cars; deliveries with autonomous drones; customer support with chat-bots; personal assistant with smart speakers ... An Artificial Intelligent agent (AI) can be trained to become expert at a task through a system of rewards and punishment, also well known as Reinforcement Learning (RL). However, since the AI will deal with human beings, it also has to follow some moral rules to accomplish any task. For example, the AI should be fair to the other agents and not destroy the environment. Moreover, the AI should not leak the privacy of users' data it processes. Those rules represent significant challenges in designing AI that we tackle in this thesis through mathematically rigorous solutions.

More precisely, we start by considering the basic RL problem modeled as a discrete Markov Decision Process. We propose three simple algorithms (UCRL-V, BUCRL and TSUCRL) using two different paradigms: Frequentist (UCRL-V) and Bayesian (BUCRL and TSUCRL). Through a unified theoretical analysis, we show that our three algorithms are near-optimal. Experiments performed confirm the superiority of our methods compared to existing techniques. Afterwards, we address the issue of fairness in the stateless version of reinforcement learning also known as multi-armed bandit. To concentrate our effort on the key challenges, we focus on two-agents multi-armed bandit. We propose a novel objective that has been shown to be connected to fairness and justice. We derive an algorithm UCRG to solve this novel objective and show theoretically its near-optimality. Next, we tackle the issue of privacy by using the recently introduced notion of Differential Privacy. We design multi-armed bandit algorithms that preserve differential-privacy. Theoretical analyses show that for the same level of privacy, our newly developed algorithms achieve better performance than existing techniques.

Keywords: Reinforcement Learning, Markov Decision Process, Multi-Armed Bandit, Multi-Agent Learning, Differential Privacy, Fairness.

Acknowledgements

This thesis would have been impossible to write without the support from many people surrounding me. This is why I want to thank a few of them here for the time they gave me.

First of all, I would like to thank my supervisor Christos Dimitratakis for the huge support you have given me. The invaluable scientific and personal pieces of advice you have been relentlessly giving me, was a key factor that helps me grow as a researcher. Even during the times were I was less productive, you have always been supportive and given me motivation to pursue the PhD. The courses that I have taught alongside you, along with the tips that you kept giving me, were also an important inspiration for the second part of my PhD research after the licentiate degree. I am really honored to be your first PhD Student and it has been a real pleasure working alongside you for these 5 years. Next, I would like to thank Debabrota Basu for the boost he has given me towards the end line of my PhD. The technical discussions that we hold time over time; the late night shifts that we spent polishing papers to meet deadlines as well as your career advice have been an immense support for the completion of my latest results. I would also like to thank Hannes Eriksson. The extensive technical discussions, your tips about lifestyle; tips about Sweden were an important aspect of my personal and professional development. You are also the key person to motivate me towards stock investments, an important part of retirement planning and wealth building. I would always be indebted to you. I would also like to thank Divya Grover for the time that you dedicated to me and the Bayesian oriented discussions. You were the final shot that transformed me into a Bayesian believer. I cannot forget to thank Emilio Jorge for the huge support, tips and tricks you have given me. I got to play real football in a real football club here in Sweden because of your encouragement. It has been a pleasure trying to unlock defenses alongside you.

I also feel honored and humbled to have Michal Valko as the opponent of my thesis. I am grateful to the grading committee members, Emilie Kaufmann, Alexandre Proutiere, Nicolò Cesa-Bianchi for having accepted reviewing my work and traveling to Gothenburg to do so. I feel grateful to Lennart Svensson for being ready to help for the defense as well as teaching me important courses during my PhD. I am grateful to Viannet Perchet who was ready to help on the defense only for administrative issues to come in the way.

I cannot write an acknowledgement without thanking Katja Hofmann, Jaroslaw Rzepecki, Sam Devlin and the many other people at MSR whose names are not mentioned here such was their huge personal and professional support since my internship with them at Microsoft; alongside the fruitful and enlightening discussions we have had afterwards. Your were a key part into growing my confidence. Your words were an essential turning

point that gave me the courage to finish my latest results even the ones unrelated to our common project. I am also grateful to Grzegorz Makosa, Max Ghenis and Duc-Hieu Tran for the excellent time I have spent with you during my internship at Google. I feel I am a better engineer because of you.

I am also thanking my co-supervisor Katerina Mitrokotsa for her huge and personal support, my examiner Devdatt Dubhashi for suggesting many useful and interesting research directions ; Graham Kemp for making me aware of an interesting conference that allowed me to disseminate my work back in my original country, Benin; Chien-Chung for accommodation help in Paris; Pablo Picazo-Sánchez, Petre Mihail Anton, Ashkan Panahi, Elena Pagin and many others for their bits of advice about life in Sweden and the various benefits at Chalmers during my early years as a PhD student. I am also sending my acknowledgement to the administration support at Chalmers such as Eva, Agneta, Fatima, Rebecca for their magic support for any issues I encountered.

I would also like to thank every member of my religious family at Fishers Creeks International as well as our "Small Group". In particular, I am thanking Elena, Paul, Simeon, Vaughan, Javier, Hamid, Raquel, Arno, Johann, Anok along with the many people whose names are not here. The discussions that we have had as well as the time that we are spending together provide me a real pleasure. In this group, I would like to give special kudos to Vaughan, Anok and Ilir for commenting on my popular science presentation. I am not forgetting to thank friends and colleagues such as Adones Rakundo, Constantin Cronrath James Wen, Lionel Houssou, Giles Dansou, Christ Felix and many others whose support have been immeasurable.

Furthermore, I feel indebted to my partner Rose Orji. Your support has been huge, literally and I could write a book about how helpful and supportive you were. But I have limited space and will keep it this short. Mum, Dad and my sisters have also been very supportive. Thank you again.

Finally, I show my gratitude to the many people whose names are not mentioned here but worked behind the scenes to help me. You are real unsung heroes.

Structure and List of publications

This thesis follows the *compilation thesis* structure commonly recommended in the technical departments of the Nordic universities.

The following manuscripts are included in the thesis.

Paper I	A. Tossou, D. Basu, and C. Dimitrakakis. Near-optimal Regret Bounds for Optimistic Reinforcement Learning using Empirical Bernstein Inequalities. *ERL - ICML'19 Workshop, Submitted to AISTATS* (2020)
Paper II	A. Tossou, C. Dimitrakakis, and D. Basu. Near-optimal Bayesian Solution For Unknown Discrete Markov Decision Process. *Submitted to STOC* (2020)
Paper III	A. Tossou et al. A Novel Individually Rational Objective In Multi-Agent Multi-Armed Bandit: Algorithms and Regret Bounds. *Submitted to AAMAS* (2020)
Paper IV	A. C. Y. Tossou and C. Dimitrakakis. "Achieving Privacy in the Adversarial Multi-Armed Bandit". *AAAI*. AAAI Press, 2017, pp. 2653–2659
Paper V	A. C. Y. Tossou and C. Dimitrakakis. "Algorithms for Differentially Private Multi-Armed Bandits". *AAAI*. AAAI Press, 2016, pp. 2087–2093

The following manuscripts have been published, but are not included in this work.

Paper VI	A. C. Y. Tossou, C. Dimitrakakis, and D. P. Dubhashi. "Thompson Sampling for Stochastic Bandits with Graph Feedback". *AAAI*. AAAI Press, 2017, pp. 2660–2666
Paper VII	A. Hossmann-Picu et al. "Synergistic user $\longleftrightarrow$ context analytics". *ICT Innovations 2015*. Vol. 399. Springer, Jan. 2016, pp. 163–172
Paper VIII	A. C. Y. Tossou and C. Dimitrakakis. "Optimal Advertisement Strategies for Small and Big Companies". *AFRICOMM*. vol. 171. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering. 2015, pp. 94–98

Paper IX A. C. Tossou and C. Dimitrakakis. On The Differential Privacy of Thompson Sampling With Gaussian Prior. *PiMLAI'18 - ICML Workshop - preprint arXiv:1806.09192* (2018)

Paper X P. Ekman et al. Learning to match. *VAMS, RecSys Workshop - arXiv preprint arXiv:1707.09678* (2017)

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Artificial Intelligence (AI) systems are now ubiquitous in many real-life applications such as self-driving cars, drone deliveries, customer support chat-bots, games, routing, robotics ... One of the most common way to train AI agents is through reinforcement learning, in which case the agents learn by trial and errors while interacting with its environment. As a result, by giving appropriate rewards and punishments, a designer can train the AI agent to become an expert at any task since the goal of AI agent will be to maximize its accumulated rewards. One of the key challenge for training this AI agent so as it can act in a way that maximizes its accumulated rewards is the exploration/exploitation dilemma. To understand this concept, let's assume that the AI agent has currently tried an action a and obtained a high reward. What should the agent do at the next round? Should the agent continue taking the action a? This is exploitation since the agent is picking an action it has a lot of information about. Or should the agent try another action in the hope that it may lead to a higher reward? This is exploration since the agent is trying an action it has few or no information about. This is a dilemma since if the agent never explores then, it may be missing on potentially larger rewards. At the same time, if the agent never exploits, it won't get enough of large rewards. In this thesis, we show how to optimally trade-off exploitation and exploration in a large class of reinforcement learning problems.

Even with the exploitation/exploration issue resolved, there are still lot of practical challenges. Indeed, the AI agent will deal with human beings. For example, we may want the AI agent to recruit candidates for a job. In this example, there are lot of moral requirements (such as not discriminating candidates based on their background) that the AI agent must guarantee. In general, whenever the AI agent is dealing with humans it

must ensure safety and fairness. If the AI agent is dealing with sensitive user data, it must ensure that its decisions do not leak the privacy of those users. Those represent constraints on the decision that the AI agent may be able to take. In this thesis, we consider the more restricted settings of stateless reinforcement learning also well-known as multi-armed bandit. In this setting, we show for a two-player system how to design an AI agent that will arrive at a rational and fair decision to both parties. We also show how the agent can arrive at a decision that will not leak privacy. This is done both for the stochastic multi-armed bandit where the rewards are generated from a fixed probability distribution as well the adversarial multi-armed bandit where the rewards can be generated arbitrarily based on a fixed memory size.

1.1 Background

In this section, we present the technical background necessary to understand the context of the contribution of this thesis. We start by introducing Markov Decision Processes, as well as their classification. We then introduce the problems of Reinforcement Learning and Multi-Armed Bandits. We finish by giving an overview of well-known equilibrium concepts for multi-agent systems.

Reinforcement learning is a sequential decision making problem, where an agent interacts with an unknown environment. At round t, the agent observes the environment, takes an action and obtains a numerical reward r_t. The agent should take actions so as to maximise total reward over time, i.e. $\sum_t r_t$. The typical model used to represent the interaction between the agent and the environment is called a Markov Decision Process (MDP), which we describe in Section 1.1.1 below. The reinforcement learning problem of an agent acting in an MDP is then described in Section 1.1.2, while we discuss learning in a multi-agent setting in Section 1.1.3.

1.1.1 Finite MDP

A finite MDP M consists of a finite state space $\mathcal{S}$, a finite action space $\mathcal{A}$, a reward distribution ν on (bounded or unbounded) real-valued rewards in $\mathcal{R}$ for all state-action pairs (s, a), and a transition kernel p such that $p(s'|s, a)$ is the probability of transiting to state s' from state s by taking an action a. At round t, a learner chooses an action $a_t \in \mathcal{A}$ according to a potentially history-dependent policy $\pi_t : (\mathcal{S} \times \mathcal{A} \times \mathcal{R})^* \times \mathcal{S} \to \Delta\mathcal{A}$ where $\Delta\mathcal{A}$ is a distribution with support on the actions, assigning probability

$$\pi_t(a_t \mid s_t, r_{t-1}, a_{t-1}, s_{t-1}, \ldots).$$

to the action a_t given the history. This grants the learner a reward $r_t(s_t, a_t)$ and transits to a state s_{t+1} according to the transition kernel p.

The optimal policy for a given MDP M requires us to define a more precise notion of optimality. In the *discounted* settings, one care about the total sum of discounted rewards where at round t, the rewards are discounted by a factor of γ^t, $\gamma < 1$. When $\gamma = 1$, this is called the *undiscounted* settings. In this thesis, we will focus on the *undiscounted* setting. A policy is called *Markov* if the action a_t only depends on the current state s_t, in which case it is not history-dependent. It is *stationary* if it plays the same state-dependent distribution at each round and it is called *deterministic* if it always assigns probability 1 to some action. The optimal policy for any MDP M is always Markov, but learning policies must necessarily be history-dependent. Given a Markov policy, the states visited form a *Markov chain*. For *undiscounted* settings, the structure of this chain is very important and finite MDPs are further subdivided into classes based on the chain structure induced by different policies. MDPs are then classified in the 5 well-known classes identified below:

1. **Recurrent or Ergodic**: Every deterministic stationary policy induces a single recurrent class (i.e. it is possible to reach any state from any other state after a finite number of steps.)

2. **Unichain** Every deterministic stationary policy induces a single recurrent class plus a possibly empty set of transient states (states that will not be visited with probably 1 after a fixed finite number of steps.)

3. **Communicating** For every pair of states s, s', there exists a deterministic stationary policy that can reach s' from s after a finite number of steps.

4. **Weakly Communicating** The set of spaces decomposed into two sets. Every state in the first set can be reached from another state in the first set after a finite number of steps under some deterministic stationary policy. The second set, possibly empty consist of states that are transient under all policies.

5. **Multichain** If there is at least one stationary policy inducing two irreductible recurrent classes.

On top of the classes described above, there is also the class of *episodic* or *finite-horizon* MDP whereby after H steps the next state distribution is the same for every policy. In particular, one can view the MDP as evolving in episodes of H steps whereby the initial state distribution at the start of each episode is the same for all policies.

In the *undiscounted* and *infinite-horizon* MDP settings a challenge is how to mathematically express the performance criteria for policies. Indeed, the total sum of rewards can go to infinity and thus no comparison could be made. One idea to solve this issue is to define the value (also gain) of a given policy π starting at s as the expected infinite-horizon average reward:

$$V(s|\pi) \triangleq \limsup_{T\to\infty} \frac{1}{T}\,\mathbb{E}\left[\sum_{t=1}^{T} r(s_t,\pi(s_t)) \mid s_1 = s\right].$$

Puterman [56] shows that there is a policy π^* whose gain, V^* is greater than that of any other policy. Also, this optimal gain is state-independent for ergodic, unichain, communicating and weakly communicating MDPs. That is $V(s|\pi^*) = V^*\ \forall s$. Furthermore, this gain satisfies the following optimality inequality for every state s:

$$h^*(s) + V^* = \max_{a\in\mathcal{A}}\left(r(s,a) + \sum_{s'\in\mathcal{S}} p(s'|s,a)h^*(s')\right). \tag{1.1.1}$$

where h^* is known as the *optimal bias* function. Intuitively, h^* quantities how good it is to start from a state compared to another one. We called *span* of h^* the quantity $\mathrm{sp}(h^*) = \max_s h^*(s) - \max_s h^*(s)$ This is an important quantity in many algorithms. Another important quantity for a finite MDP is its diameter.

Definition 1.1.1 (Diameter of a finite MDP)**.** The diameter D of an MDP M is defined as the minimum number of rounds needed to go from one state s and reach any other state s' while acting using some deterministic policy. Formally,

$$D(M) = \max_{s\neq s', s, s'\in\mathcal{S}} \min_{\pi:\mathcal{S}\to\mathcal{A}} T(s'|s,\pi).$$

where $T(s'|s,\pi)$ is the expected number of rounds it takes to reach state s' from s using policy π.

For communicating MDPs, by definition this quantity is finite.

1.1.2 Reinforcement Learning and Multi-Armed Bandit

In reinforcement learning, the MDP M is assumed to be *unknown*. In this thesis we assume that the agent always observes the state of the MDP s_t before it takes its action a_t, that it observes the reward r_t immediately afterwards. We focus on the case where the rewards r are bounded in $[0,1]$.

In the undiscounted setting, to compare different algorithm for reinforcement learning, we can look at the amount of reward obtained by that algorithm compared to what an optimal policy would have obtained after the same amount of time. This is called the regret, defined below as:

$$\text{Regret}(T) \triangleq \sum_{t=1}^{T} \left(V^* - r(s_t, a_t)\right).$$

The goal in reinforcement learning is thus equivalent to minimizing this regret. This regret is commonly known as the (frequentist) regret. There is another notion of regret called Bayesian regret commonly used by Bayesian algorithms. To understand Bayesian regret, let's first observe that since the true MDP is unknown, Bayesian methods start with a prior distribution for the unknown MDP and compute a posterior distribution with more data collected using Bayes theorem. The Bayesian regret can be viewed as the expected frequentist regret under the assumption that the true unknown MDP comes from the prior. More precisely, the Bayesian setting assumes that the true MDP is in some set $\mathcal{M} = \{ M_\theta \mid \theta \in \Theta \}$ parameterized by θ, over which a prior probability distribution $\mathbb{P}$ is defined. And the Bayesian Regret for a policy π is defined as:

$$B\text{Regret}(T) = \int_\Theta \mathbb{E}^\pi_{M_\theta} \text{Regret}(T) \, \mathrm{d}\, \mathbb{P}(\theta). \tag{1.1.2}$$

Unless otherwise specified, in this thesis we used regret to mean the frequentist regret. The minimax regret is a worst-case problem independent regret. To find it, one designs a worst-case MDP and shows that under this MDP, no algorithms can obtain a regret smaller than a specific value. More precisely, the minimax regret is defined by:

$$\min_{\pi \in \Pi} \max_{M \in \mathcal{M}} \mathbb{E}^\pi_M \text{Regret}(T).$$

When the MDP consists of a single state, the corresponding reinforcement learning problem is known as the *stochastic multi-armed bandit* problem.

1.1.3 Multi-Agent Learning

The extension of MDP to multiple players is known as *stochastic game* or *Markov game*. A finite *stochastic game* consists of a finite set of players $\mathcal{N}$, a finite state space $\mathcal{S}$, a finite set of joint-actions $\mathcal{A} = \prod_{i=1}^{|\mathcal{N}|} \mathcal{A}^i$ with $\mathcal{A}^i$ the set of actions available for player i. At each round t, from the state s_t the players simultaneously take actions $a_t = a_t^1, \ldots a_t^{|\mathcal{N}|}$ where a_t^i is the action taken by player i. Each player i then receives a numerical reward

$r^i(s_t, a_t)$ sampled from some probability distribution depending only on the current state s_t and the joint action a_t. Then, the system transits to a new state s_{t+1} according to a probability distribution that depends only on (s_t, a_t). We say that the rewards are *deterministic* if given a state s_t, a joint-action a_t and a player i, the reward $r^i(s_t, a_t)$ is always the same value. If this is not the case, then we say that the rewards are *stochastic*.

Stochastic games with a single-state are known as *repeated (general sum) games*. If the rewards are deterministic and always sum to zero at each round t, the end of the game, then the game is known as being *zero-sum*. If the game only lasts for 1 round, it is known as *single-stage* or *one-shot* game.

Most of the literature relating to games consider only deterministic rewards. Similarly to MDPs, players act through policies. When the policy is *deterministic stationary* this is known as a *strategy* in the game theory literature. A strategy profile is a list of $|\mathcal{N}|$ strategies, one for each player. Given a strategy profile, each player can compute their expected long-term rewards. Just as in MDPs, this is called the value. Each player would like to maximize their own value.

While from the point of view of one player, maximising total utility might be a reasonable goal, as players can reason about how other players might act, there may not exist a strategy that is optimal in the single-agent sense. For that reason, alternative solutions concepts have been developed, such as the well-known Nash Equilibrium [48]. A strategy profile is said to be a Nash Equilibrium if no players can change its strategy in the profile and increase its value as far as the other players do not change their strategies in the profile.

Another important class of solutions concept is that of maximin. Informally, the maximin is the largest value that a player can guarantee regardless of the policies of the other players, i.e. where it assumes that all the other players are acting as a coalition against him in a zero-sum game. Similarly, the minimax is the smallest value the other players can force a specific player to have regardless of its policy.

In this thesis, we are instead targeting another solution concept, which is more closely related to reinforcement learning and multi-armed bandit problems, the Egalitarian Bargaining Solution whereby through cooperation players can increase their value.

1.2 Thesis outline

In the remainder of this chapter, we detail our main contributions, then we review the related literature. We finish this chapter with concluding remarks. In Chapter 2, we present in details our article that introduced a novel optimistic algorithm, UCRL-V for

communicating Markov Decision Process. Chapter 3 extends the results from Chapter 2 to the Bayesian viewpoint. In particular, we present two algorithms BUCRL and TSUCRL that respectively uses the quantile and the sampled order statistics of the posterior distribution to derive confidence bounds to use when solving the MDP. Chapter 4 detailed our work on multi-agent multi-armed bandit and proposes a novel objective: Egalitarian Bargaining Solution. We also present an algorithm that can achieve this objective.

1.3 Contributions

Our main contributions can be summarized as follows:

- We derive a computationally efficient optimistic algorithm, UCRL-V, for communicating MDP that achieves $\tilde{\mathcal{O}}(\sqrt{DSAT})$[1] regret closing a gap in the literature. Our theoretical analysis is generic and can be applied to other algorithms. Experiments conducted confirmed our theoretical results ([66] and re-printed in Chapter 2).

- We derive the first computationally efficient Bayesian algorithms, BUCRL and TSUCRL, for communicating MDP that achieve the optimal regret $\tilde{\mathcal{O}}(\sqrt{DSAT})$ up to logarithmic factors ([67] and re-printed in Chapter 3).

- We propose a novel objective (the Egalitarian Bargaining Solution) to aim for in multi-agent multi-armed bandit. We also proposed an equivalent notion of individually rational regret. We propose an algorithm UCRG that achieves the near-optimal regret in this settings. We can conclude that the achieved regret is near-optimal since we demonstrated a matching lower bound ([68] and re-printed in Chapter 4).

- We design differential private multi-armed bandit algorithm that improved the regret suffered for ensuring a fixed ϵ privacy loss from $\tilde{\mathcal{O}}\big(T^{2/3}/\epsilon\big)$ to $\tilde{\mathcal{O}}\Big(\sqrt{T}/\epsilon\Big)$ (Chapters 5 and 6).

1.4 Related work

In this section, we review the literature which we categorized as related to *optimistic reinforcement learning*, *posterior sampling for reinforcement learning* and *multi-agent learning*.

[1] $\tilde{\mathcal{O}}$ is used to hide log factors

Optimistic algorithms Jaksch, Ortner, and Auer [31] were one of the first to propose an algorithm for the general communicating MDP with undiscounted reward. They introduce an algorithm named UCRL2 and demonstrate an upper bound of $\mathcal{O}(DS\sqrt{TA})$ on the regret. It is important to note that UCRL2 never uses any information about the true diameter D. UCRL2 works by constructing adaptive episodes. The criteria used to construct a new episode is named *doubling trick*. Within an episode, UCRL2 builds a set of statistically plausible MDPs and finds within that set, an MDP and its optimal policy (called *optimistic policy*) such that the value of that optimistic policy is close to the largest value attainable for any other MDP in the set. This step is called *optimism* and is the reason why the method is called *optimism in the face of uncertainty*. The algorithm used to find this optimistic policy is called *extended value iteration*. Jaksch, Ortner, and Auer [31] show that their result can be improved since they prove a lower bound of $\Omega(\sqrt{DSAT})$ on the regret. They also derived a nice and intuitive analysis that turned out to become the backbone of future analysis.

Filippi, Cappé, and Garivier [24] derive an algorithm named KL-UCRL that follows the same structure as UCRL2. Their main modification comes from how the set of statistically plausible MDPs is constructed. More precisely, instead of relying on Weissman to create a bound on the transitions they used a confidence interval based on KL-divergence. Because of this modification, KL-UCRL modifies how to find the optimistic policy. In particular, to find the MDP whose optimal policy leads to an *optimistic* policy, KL-UCRL solves a convex maximization over a ball using Newton-Raphson algorithm. Theoretically, KL-UCRL does not improve on the upper bound of the regret compared to UCRL2. However, Filippi, Cappé, and Garivier [24] found that in practice, KL-UCRL performs much better than UCRL2.

Talebi and Maillard [64] proposes an improved regret bound for KL-UCRL in the more restricted setting of ergodic MDPs. They show a regret bound of $\mathcal{O}(\sqrt{ST\sum_{s,a}\boldsymbol{V}^{*}_{s,a}} + D\sqrt{T})$ where $\boldsymbol{V}^{*}_{s,a}$ is the variance of the optimal bias function with respect to the next state distribution when following action a in state s. Their analysis suggest why KL-UCRL is much better in practice than UCRL2. Indeed, for some MDPs this variance term is much lower than D^2SA. However, in the worst-case it could reach D^2SA and thus the result of Talebi and Maillard [64] does not provide any improvement.

There have been also some papers that attempted to improve the dependency on D and instead have a dependency on $\mathrm{sp}(h^*)$ which is always a much smaller quantity. Although the lower bound of Jaksch, Ortner, and Auer [31] shows a dependency on D, it uses a worst-case MDP where $\mathrm{sp}(h^*)$ and D are within a constant factor of each other. This has generated a lot of discussion [25] in the literature to know whether or not the

"true" lower bound should depend $\text{sp}(h^*)$ rather than D. The intuition was "While the diameter D quantifies the number of steps needed to "recover" from a bad state in the worst case, the actual regret incurred while "recovering" is related to the difference in potential reward between "bad" and "good" states, which is accurately measured by the span (i.e., the range) $\text{sp}(h^*)$ of the optimal bias function. While the diameter is an upper bound on the optimal bias span, it could be arbitrarily larger (e.g., weakly-communicating MDPs may have finite span and infinite diameter) thus suggesting that algorithms whose regret scales with the span may perform significantly better."

Bartlett and Tewari [9] proposed the *Regal* algorithm with $\mathcal{O}\Big(\text{sp}(h^*)S\sqrt{TA}\Big)$. However, not only their algorithm is not efficiently implementable it requires the knowledge of the true $\text{sp}(h^*)$. Although Bartlett and Tewari [9] presented two more algorithms Regal.C and Regal.D that only need an upper bound (such as D) on $\text{sp}(h^*)$, their algorithm remains not efficiently implementable. This is because instead of finding the policy maximizing the gain in the set of statistically plausible MDPs (something that can be done efficiently using *extended value iteration*), Bartlett and Tewari [9] use an additional regularization term and propose to find a policy that maximizes the sum of the gain with this regularized term. This leads to a non-convex optimization problem which is NP-hard to solve.

Fruit et al. [25] aims to pick up where Bartlett and Tewari [9] left and find an implementable algorithm that scale with the span of the optimal bias function. They relaxed the optimization in Regal.C and were able to provide an efficient algorithm to solve this relaxation. Their algorithm named SCAL enjoys $\mathcal{O}\Big(\text{sp}(h^*)\sqrt{SA\Gamma T}\Big)$ where Γ is the maximum number of possible next states. Since their regret depends on $\text{sp}(h^*)$ and Γ, it provides a slight improvement over the regret of UCRL2. However, in the worst case both regret are identical and still far from the achievable lower bound on the regret.

Azar, Osband, and Munos [7] focus on the more restricted setting of episodic MDP with known episodic length H. However, they took a completely different approach compared to UCRL2. Instead of building a set of statistically plausible set for the MDP, they built a plausible set directly around the optimal value function (or gain). They show that their algorithm named UCBVI can achieve $\mathcal{O}\Big(\sqrt{HSAT}\Big)$. However, this near-optimal regret is only achieved when when $T > H^3S^3A$.

Following the results of [7], more recent works have attempted to obtain the achievable lower bound on the regret. For example, Efroni et al. [22] and Simchowitz and Jamieson [60] propose algorithms achieving $\mathcal{O}(\sqrt{HSTA})$ for episodic MDP with known episode length H. Zhang and Ji [71] show an algorithm that can achieve $\mathcal{O}(\sqrt{DSAT})$ for weakly communicating MDP. However, their algorithm is not efficiently implementable as it depends on a NP-hard non-convex optimization. In addition, their regret bound only

holds when T is larger than a polynomial function in $(D, S, A, \log T)$ with degree at least 2 for the terms D, S.

All the previously mentioned algorithm are model-based in the sense that they maintain an explicit model (that is the rewards and transitions) of the MDP and uses this model internally to obtain the policy to play. There has also been some works for model-free reinforcement learning. In particular, Jin et al. [32] shows that Q-Learning with UCB-style exploration can achieve $\mathcal{O}(\sqrt{H^3SAT})$ for episodic MDPs with known H.

Bayesian algorithms There have been fewer results using Bayesian based algorithms to solve undiscounted MDPs. Many of the Bayesian algorithms such as [4] only work for discounted MDP and provide PAC (Probably Approximately Correct) bounds and not regret bounds. Translating the PAC bounds to regret bound would lead in the best case to $\mathcal{O}\big(T^{2/3}\big)$ regret as explained in [31].

One of the first regret analysis for undiscounted MDP using a Bayesian algorithm is due to Osband, Russo, and Van Roy [50]. They focus on episodic MDP with known episode length H. Their algorithm named PSRL start with a prior over MDP and compute the Posterior using Bayes' theorem. At the beginning of each episode, PSRL samples an MDP from the posterior, finds the optimal policy for this sampled MDP and follows that policy for the remainder of the episode. Osband, Russo, and Van Roy [50] show that this computationally efficient PSRL enjoys $\mathcal{O}\Big(HS\sqrt{TA}\Big)$ Bayesian regret. Since Osband, Russo, and Van Roy [50] paper, there have been a lot of failed attempts to extend PSRL to infinite-horizon MDP such as [1, 2]. Indeed, [1] proposes a version of PSRL for infinite-horizon with regret analysis, however Osband and Van Roy [51] pointed to a mistake in their analysis. Similarly, Agrawal and Jia [2] proposes an analysis of an optimistic version of PSRL for infinite-horizon MDP. However, Fruit et al. [25] and Zhang and Ji [71] have pointed to mistakes into their analysis. The main technical challenges into adapting PSRL for infinite-horizon problems stem from the fact that for infinite-horizon one usually need to create adaptive data-dependent episodes which is problematic for PSRL [51].

To solve this problem, Kim [36] proposes to sample a new MDP from the posterior at each round, find its optimal policy and follow it. This process is computationally expensive and Kim [36] could only provide an asymptotic regret (that is when the total number of rounds tend to infinity). Ouyang et al. [54] were able to provide an extension to PSRL which they named TSDE achieving $\mathcal{O}\Big(HS\sqrt{TA}\Big)$ Bayesian regret for weakly communicating MDP. The main contribution in TSDE is how the dynamic episodes are constructed. TSDE combines the doubling trick with another criterion that limits how

large the length of an episode can grow to.

Multi-Agent Reinforcement Learning/Multi-Armed Bandit There is a huge literature in multi-agent learning. In this review, we will focus on reviewing cooperative multi-agent learning which is where we made our main contribution.

There is a growing interest in multi-agent multi-armed bandit. Many of the works [3, 5, 6, 10, 13, 18, 23, 27, 28, 35, 37–39, 44, 45, 49, 58, 59, 63, 69] have focused on maximizing the *sum of rewards among players* also sometimes known as a type of social welfare function. The main arguments for using the *sum of rewards* come from the assumption that there exists a centralized designer who want the system of individual agents to converge to a globally good solution taken as the *sum of rewards*. However, the *sum of rewards* may not make sense if the agents are rational since it is possible for an agent to obtain lower than what it could have obtained without cooperation regardless of the strategies of the other agents Brafman and Tennenholtz [16]. More precisely, any agent can always guarantee its maximin value whereas the value an agent can obtain when maximizing for the *sum of rewards* may be lower than the maximin. As a result, the outcome is not fair to the agent that will receive lower than the maximin.

Another vast amount of works have focused on aiming for some single-stage Equilibrium such as single-stage Nash Equilibrium or single-stage Correlated Equilibrium [8, 14, 15, 19, 61]. The majority of research in this line of works have considered deterministic rewards and some asymptotic convergence analysis.

Others consider discounted rewards settings [26, 29, 30, 40, 41, 43, 72] using some variants of Q-learning. Although, it may be possible to easily extend those algorithms to stochastic rewards, they consider discounted rewards and only asymptotic convergence analysis are provided.

Some work in this line [8, 14] provides a notion of "no-regret". We would like to note that their notion of regret is not comparable to ours. In their settings, the rewards are still deterministic. However, they provide algorithms that need multiple rounds to converge to an equilibrium. Their notion of "regret" relates to how much they lose while converging.

Furthermore, aiming for single-stage equilibrium when the game is repeated is problematic [16]. For example, consider the iterated prisoner dilemma game. It is a game between two players each having two actions: Cooperate or Defect. The game is constructed such that the only single-stage Nash Equilibrium is for both players to Defect with each agent receiving -2 as rewards. However, both player can Cooperate and receive each -1 as reward. In other words, by cooperating both players can receive much higher than they would receive in any single-stage equilibrium. Acknowledging the issues with

single-stage equilibrium, Powers, Shoham, and Vu [55] propose a new criterion which in cooperative settings implies *Individual Rationality*: the average reward is Pareto efficient and individually not below the maximin value. To this effect, Powers, Shoham, and Vu [55] propose the PCM(A) algorithm, that maximizes the *sum of rewards* among the strategies that are pareto-optimal and not below the maximin value. However, PCM(A) only consider the case of deterministic rewards. Also, there is no justification as to why one should aim for maximizing the *sum of rewards* among the strategies not below the maximin value.

In fact, this remark is a much larger issue. Indeed, various *folk theorems* [53] suggest that for infinite-horizon undiscounted reward, every outcome that is feasible and individually rational can be realized as a Nash Equilibrium of the repeated game. In short, the set of individually rational strategies may be infinite. So the main question is which one should we aim for and why? To concentrate on this question and capture the fundamental challenges, we focus on the 2-agent setting.

For 2 agents, this question has received a lot of attention and known as *Nash Bargaining Problem* [47]. In a Nash Bargaining problem, the two agents have a disagreement point (in our case the point formed using the maximin of both agents) and neither player want to receive lower than its corresponding value at the disagreement point. The objective in Nash Bargaining problem is which agreement should both player reach? A lot of solutions (Nash [47], Kalai–Smorodinsky [34], Egalitarian [33], Utilitarian [65] Bargaining Solutions) have been proposed to solve this *Nash Bargaining Problem* that are all based on formal Mathematical axioms that a solution should possess. We pick the Egalitarian Bargaining Solution (EBS) since, as opposed to the other solutions, it has been shown to be connected to some concepts that are sought-after for autonomous agents in our society today. These concepts include *fairness*, *equality* and Rawls (1971) theory of justice for human society. EBS also enjoys strong mathematical properties. On top of the individual rational criterion, it also satisfies independence of irrelevant alternatives (i.e. eliminating choices that were irrelevant does not change the choices of the agents), individual monotonicity (If a player has better options in one game compared to another game, then that player should get a weakly-better value in the game with better options) and (importantly) uniqueness.

There has been previous works that also consider using solutions to Nash Bargaining Problem as an objective in the 2-agent repeated game. Munoz de Cote and Littman [46] provides an algorithm to find the EBS for general-sum repeated stochastic games with deterministic rewards and known transitions. As a result, they don't provide a regret analysis caused by uncertainty in the rewards. Furthermore, even when applied to a

known multi-armed bandit with deterministic rewards, their algorithm implies finding an approximate solution using a (binary) search in the space of policies. In constract, in our thesis, we derived an exact solution with a direct and simple formula. [42] provides a solution for general-sum repeated games using the Nash Bargaining Solution (NBS). However, here again they consider deterministic rewards.

Since our proposed EBS is individually rational, it can be realized as a Nash Equilibrium of the repeated game. Brafman and Tennenholtz [16] have also proposed an algorithm that converges to a Nash Equilibrium of the repeated game. Their solution maximizes the surplus above the maximin of the players. However, they also consider deterministic rewards. A similar idea could be used even when the rewards are stochastic. That is play each join-action for m rounds and then use the empirical observed game to compute the desired policy. This heuristic is well-known as Explore-Then-Commit (ETC) in the multi-armed bandit literature. We perform experimental analysis against this strategy with m taken as the minimax optimal value (the value minimizing the regret in the worst-case game) and we show that our proposed algorithm outperforms ETC.

Crandall and Goodrich [20], Qiao et al. [57], and Stimpson and Goodrich [62] propose algorithms with the goal of converging to a NE of the repeated games, more precisely the NBS. All of them consider deterministic rewards and only provide asymptotic convergence to the NBS. Furthermore, Crandall and Goodrich [20] and Qiao et al. [57] consider the discounted rewards settings. While Stimpson and Goodrich [62] only argue that there exists some parameters of their algorithm that will make it more likely to converge to the NBS. They do not give the actual value of those parameters and note that convergence to the NBS would be slow.

Brafman and Tennenholtz [17] and Wei, Hong, and Lu [70] tackle online learning for a generalization of repeated games called *stochastic games*. However, they consider zero-sum games where the sum of the rewards of both players for any joint-action is always 0. In our case, we look at the general sum case where no such restrictions are placed on the rewards.

Our work is also related to multi-objective multi-armed bandit [21] by considering the joint-actions as arms controlled by a single-player. Typical work consider on multi-objective multi-armed bandit tries to find any solution that minimizes the distance between the Pareto frontier. However, not all Pareto efficient solutions are acceptable as illustrated by Example 4.2.1 in our paper Tossou et al. [68]. Instead, our work show that a specific Pareto efficient (the EBS) is more desirable.

1.5 Concluding Remarks and Future Directions

In this thesis, we presented three algorithms UCRL-V (based on optimism principle) and BUCRL, TSUCRL (based on posterior sampling) that solve communicating MDPs and achieve the optimal regret bound up to logarithmic factors. The main takeaway from our analysis is the importance of using variance based algorithms for online decision problems.

Bayesian Algorithms vs Pure Optimism Bayesian algorithms naturally include implicitly variance which may explain their historial outperformance compared to optimistic algorithms. In this thesis, we show that using variance confidence bounds, optimistic algorithms can match their Bayesian counterpart up to constant factor. We thereby disprove one commonly believed conjecture in the field [52] in which optimistic algorithms are thought to always be inferior. As a conclusion, theoretically there is no difference in what can be achieved with Bayesian methods compared to pure optimistic methods. And this remains true even when more information about the distribution of the MDP is known. However, Bayesian methods remain more attractive from a practical point of view. In particular, it is much easier to include prior knowledge about the MDP into Bayesian methods compared to optimistic ones. There is a wide variety of conjugate priors for which an analytical solution to the posterior is available. Even when the prior is very complex one can use a variety of approximation techniques such as variational inference to obtain a good approximation of the posterior. Availability of Python library Pyro [12] facilitate this even more.

We also proposed a solution for one the main dilemma in multi-agent system which is the criteria one should aim for. We suggested the use an Egalitarian Bargaining Solution and demonstrated its superiority against competing solution concepts. We also derived near-optimal algorithms in this settings. Finally, we demonstrate how to simultaneously preserve differential privacy and optimize regret in multi-armed bandits problems.

1.5.1 Future Directions

Natural Extension of BUCRL to non-Bernoulli rewards Currently, BUCRL works for non-Bernoulli rewards by performing a Bernoulli trials on the observed rewards. We believe this is unnecessary and that one can directly use the observed rewards. The main intuition behind this idea is that Bernoulli distribution is the distribution with maximal variance among all bounded distribution in $[0, 1]$ [11]. As a result, the corresponding Binomial quantile should be larger than the corresponding quantile for any other distribution.

Extension of our results for UCRL-V and BUCRL from communicating MDP to weakly communicating MDP We believe this should be a trivial extension.

Extension of our results for UCRL-V, BUCRL and TSUCRL from communicating MDP to MultiChain MDP This will not be a trivial extension. One of the key dilemma in multi-chain is that some states may be unreachable by our learning algorithm and yet be reachable by the optimal policy. Consider an MDP with a good state s and a bad state s' where it is possible to transition from s to s' but not the other way around. Then, if a learning agent reaches state s' from s (for example because of exploration), there is no way it can achieve the good value of s and will thus suffer a large (linear) regret. We believe one of the effective way to solve this issue is to modify the oracle against which comparison is made for the regret.

Diameter D or Span $\mathrm{sp}(h^*)$ Some authors [25] believe that the lower bound should depend on the much smaller $\mathrm{sp}(h^*)$ and not D and have attempted to obtain bounds scaling with $\mathrm{sp}(h^*)$. We believe that our algorithm may scale with $\mathrm{sp}(h^*)$ and not D. The only place D is enforced in our proof technique is to bound the value $\max_s u_i(s) - \min_s u_i(s)$ obtained after the extended value iteration. We believe that once each state-action is played sufficiently enough time, i.e $\mathcal{O}(\log T)$ times, then $\max_s u_i(s) - \min_s u_i(s)$ should be very close to $\mathrm{sp}(h^*)$. To play state-action that many times; we may loose up to $DSA \log T$. As a result, an upper bound of $\tilde{\mathcal{O}}\Big(\sqrt{\mathrm{sp}(h^*)SAT} + DSA \log T\Big)$ may be possible to prove.

Near-optimal pure Thompson sampling based approach Currently, our Bayesian algorithm BUCRL avoid sampling from the posterior to ensure optimism. Instead, it uses the quantile of the posterior. Whereas TSUCRL draw samples but used them to estimate quantiles. A key challenge is whether one can use the samples in a more *traditional* way similar to PSRL and still guarantee near-optimality. This is a significant open question and the key challenge is how to ensure optimism. We believe multiple samples would still be needed to avoid the policies to oscillate too often.

Extension to Function Approximators Any interesting direction would be to extend our results to function approximator and in particular to linear or quadratic function approximators.

Extending UCRG to multiple players We believe this should be a trivial extension if time complexity is not an issue. However, if one desire to obtain an algorithm with time that scale sub-linearly with the number of players, that may be more challenging.

Extend UCRG to stateful games An interesting extension of UCRG is to the full Markov games settings with states. Our results about the Existence of the EBS and its form; and the existence of two deterministic policies that can achieved the EBS value extend naturally to Markov games. The challenge is how to design a convergent algorithm that can minimize regret.

Using variance based confidence interval for UCRG We believe using Bernstein based confidence bounds should lead to improve regret $\tilde{\mathcal{O}}(\sqrt{T})$ in many games except the worst case games. This could be an important practical extension.

References

[1] Y. Abbasi-Yadkori and C. Szepesvári. "Bayesian Optimal Control of Smoothly Parameterized Systems." *UAI*. 2015, pp. 1–11.

[2] S. Agrawal and R. Jia. "Optimistic posterior sampling for reinforcement learning: worst-case regret bounds". *Advances in Neural Information Processing Systems*. 2017, pp. 1184–1194.

[3] A. Anandkumar et al. Distributed algorithms for learning and cognitive medium access with logarithmic regret. *IEEE Journal on Selected Areas in Communications* **29**.4 (2011), 731–745.

[4] J. Asmuth et al. "A Bayesian sampling approach to exploration in reinforcement learning". *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*. AUAI Press. 2009, pp. 19–26.

[5] O. Avner and S. Mannor. "Concurrent bandits and cognitive radio networks". *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer. 2014, pp. 66–81.

[6] O. Avner and S. Mannor. "Multi-user lax communications: a multi-armed bandit approach". *IEEE INFOCOM 2016-The 35th Annual IEEE International Conference on Computer Communications*. IEEE. 2016, pp. 1–9.

[7] M. G. Azar, I. Osband, and R. Munos. Minimax regret bounds for reinforcement learning. *arXiv preprint arXiv:1703.05449* (2017).

[8] B. Banerjee and J. Peng. "Performance bounded reinforcement learning in strategic interactions". *AAAI*. Vol. 4. 2004, pp. 2–7.

[9] P. L. Bartlett and A. Tewari. "REGAL: A Regularization Based Algorithm for Reinforcement Learning in Weakly Communicating MDPs". *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*. UAI '09. AUAI Press, 2009, pp. 35–42.

[10] L. Besson and E. Kaufmann. "Multi-Player Bandits Revisited". *Algorithmic Learning Theory*. 2018, pp. 56–92.

[11] R. Bhatia and C. Davis. A better bound on the variance. *The American Mathematical Monthly* **107**.4 (2000), 353–357.

[12] E. Bingham et al. Pyro: Deep Universal Probabilistic Programming. *arXiv preprint arXiv:1810.09538* (2018).

[13] I. Bistritz and A. Leshem. "Distributed multi-player bandits-a game of thrones approach". *Advances in Neural Information Processing Systems*. 2018, pp. 7222–7232.

[14] M. Bowling. "Convergence and no-regret in multiagent learning". *Advances in neural information processing systems*. 2005, pp. 209–216.

[15] M. Bowling and M. Veloso. "Convergence of Gradient Dynamics with a Variable Learning Rate". *In Proceedings of the Eighteenth International Conference on Machine Learning*. 2001, pp. 27–34.

[16] R. I. Brafman and M. Tennenholtz. "Efficient learning equilibrium". *Advances in Neural Information Processing Systems*. 2003, pp. 1635–1642.

[17] R. I. Brafman and M. Tennenholtz. R-max-a general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research* **3**.Oct (2002), 213–231.

[18] N. Cesa-Bianchi, C. Gentile, and Y. Mansour. Delay and cooperation in nonstochastic bandits. *The Journal of Machine Learning Research* **20**.1 (2019), 613–650.

[19] V. Conitzer and T. Sandholm. AWESOME: A general multiagent learning algorithm that converges in self-play and learns a best response against stationary opponents. *Machine Learning* **67**.1-2 (2007), 23–43.

[20] J. W. Crandall and M. A. Goodrich. Learning to compete, coordinate, and cooperate in repeated games using reinforcement learning. *Machine Learning* **82**.3 (2011), 281–314.

[21] M. M. Drugan and A. Nowe. Designing multi-objective multi-armed bandits algorithms: A study. *learning* **8** (2013), 9.

[22] Y. Efroni et al. Tight Regret Bounds for Model-Based Reinforcement Learning with Greedy Policies. *arXiv preprint arXiv:1905.11527* (2019).

[23] N. Evirgen and A. Kose. "The effect of communication on noncooperative multi-player multi-armed bandit problems". *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE. 2017, pp. 331–336.

[24] S. Filippi, O. Cappé, and A. Garivier. "Optimism in reinforcement learning and Kullback-Leibler divergence". *Communication, Control, and Computing (Allerton), 2010 48th Annual Allerton Conference on*. IEEE. 2010, pp. 115–122.

[25] R. Fruit et al. “Efficient Bias-Span-Constrained Exploration-Exploitation in Reinforcement Learning”. *International Conference on Machine Learning*. 2018, pp. 1573–1581.

[26] A. Greenwald and K. Hall. “Correlated-Q Learning”. *Proceedings of the Twentieth International Conference on International Conference on Machine Learning*. ICML’03. Washington, DC, USA: AAAI Press, 2003, pp. 242–249. ISBN: 1-57735-189-4. URL: http://dl.acm.org/citation.cfm?id=3041838.3041869.

[27] A. Heliou, J. Cohen, and P. Mertikopoulos. “Learning with bandit feedback in potential games”. *Advances in Neural Information Processing Systems*. 2017, pp. 6369–6378.

[28] E. Hillel et al. “Distributed exploration in multi-armed bandits”. *Advances in Neural Information Processing Systems*. 2013, pp. 854–862.

[29] J. Hu and M. P. Wellman. Nash Q-learning for general-sum stochastic games. *Journal of machine learning research* **4**.Nov (2003), 1039–1069.

[30] J. Hu, M. P. Wellman, et al. “Multiagent reinforcement learning: theoretical framework and an algorithm.” *ICML*. Vol. 98. Citeseer. 1998, pp. 242–250.

[31] T. Jaksch, R. Ortner, and P. Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research* **11**.Apr (2010), 1563–1600.

[32] C. Jin et al. “Is q-learning provably efficient?” *Advances in Neural Information Processing Systems*. 2018, pp. 4863–4873.

[33] E. Kalai. Proportional solutions to bargaining situations: interpersonal utility comparisons. *Econometrica: Journal of the Econometric Society* (1977), 1623–1630.

[34] E. Kalai, M. Smorodinsky, et al. Other solutions to Nash’s bargaining problem. *Econometrica* **43**.3 (1975), 513–518.

[35] D. Kalathil, N. Nayyar, and R. Jain. Decentralized learning for multiplayer multiarmed bandits. *IEEE Transactions on Information Theory* **60**.4 (2014), 2331–2345.

[36] M. J. Kim. Thompson sampling for stochastic control: The finite parameter case. *IEEE Transactions on Automatic Control* **62**.12 (2017), 6415–6422.

[37] N. Korda, B. Szörényi, and L. Shuai. “Distributed clustering of linear bandits in peer to peer networks”. *Journal of machine learning research workshop and conference proceedings*. Vol. 48. International Machine Learning Societ. 2016, pp. 1301–1309.

[38] L. Lai, H. Jiang, and H. V. Poor. “Medium access in cognitive radio networks: A competitive multi-armed bandit framework”. *2008 42nd Asilomar Conference on Signals, Systems and Computers*. IEEE. 2008, pp. 98–102.

[39] P. Landgren, V. Srivastava, and N. E. Leonard. "Distributed cooperative decision-making in multiarmed bandits: Frequentist and Bayesian algorithms". *2016 IEEE 55th Conference on Decision and Control (CDC)*. IEEE. 2016, pp. 167–172.

[40] M. L. Littman. "Friend-or-foe Q-learning in general-sum games". *ICML*. Vol. 1. 2001, pp. 322–328.

[41] M. L. Littman. "Markov games as a framework for multi-agent reinforcement learning". *Machine learning proceedings 1994*. Elsevier, 1994, pp. 157–163.

[42] M. L. Littman and P. Stone. "A Polynomial-time Nash Equilibrium Algorithm for Repeated Games". *Proceedings of the 4th ACM Conference on Electronic Commerce*. EC '03. San Diego, CA, USA: ACM, 2003, pp. 48–54. ISBN: 1-58113-679-X. DOI: 10.1145/779928.779935. URL: http://doi.acm.org/10.1145/779928.779935.

[43] M. L. Littman and C. Szepesvári. "A generalized reinforcement-learning model: Convergence and applications". *ICML*. Vol. 96. 1996, pp. 310–318.

[44] H. Liu, K. Liu, and Q. Zhao. Learning in a changing world: Restless multiarmed bandit with unknown dynamics. *IEEE Transactions on Information Theory* **59**.3 (2012), 1902–1916.

[45] K. Liu and Q. Zhao. Distributed learning in multi-armed bandit with multiple players. *IEEE Transactions on Signal Processing* **58**.11 (2010), 5667–5681.

[46] E. Munoz de Cote and M. L. Littman. "A Polynomial-time Nash Equilibrium Algorithm for Repeated Stochastic Games". *Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence (UAI)*. Corvallis, Oregon: AUAI Press, 2008, pp. 419–426. URL: http://uai2008.cs.helsinki.fi/UAI_camera_ready/munoz.pdf.

[47] J. F. Nash Jr. The bargaining problem. *Econometrica: Journal of the Econometric Society* (1950), 155–162.

[48] J. F. Nash et al. Equilibrium points in n-person games. *Proceedings of the national academy of sciences* **36**.1 (1950), 48–49.

[49] N. Nayyar, D. Kalathil, and R. Jain. On regret-optimal learning in decentralized multiplayer multiarmed bandits. *IEEE Transactions on Control of Network Systems* **5**.1 (2016), 597–606.

[50] I. Osband, D. Russo, and B. Van Roy. "(More) efficient reinforcement learning via posterior sampling". *Advances in Neural Information Processing Systems*. 2013, pp. 3003–3011.

[51] I. Osband and B. Van Roy. Posterior sampling for reinforcement learning without episodes. *arXiv preprint arXiv:1608.02731* (2016).

[52] I. Osband and B. Van Roy. "Why is Posterior Sampling Better than Optimism for Reinforcement Learning?" *Proceedings of the 34th International Conference on Machine Learning*. Ed. by D. Precup and Y. W. Teh. Vol. 70. Proceedings of Machine Learning Research. International Convention Centre, Sydney, Australia: PMLR, June 2017, pp. 2701–2710. URL: `http://proceedings.mlr.press/v70/osband17a.html`.

[53] M. J. Osborne and A. Rubinstein. *A course in game theory*. 1994.

[54] Y. Ouyang et al. "Learning unknown markov decision processes: A thompson sampling approach". *Advances in Neural Information Processing Systems*. 2017, pp. 1333–1342.

[55] R. Powers, Y. Shoham, and T. Vu. A general criterion and an algorithmic framework for learning in multi-agent systems. *Machine Learning* **67**.1 (May 2007), 45–76.

[56] M. L. Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

[57] H. Qiao et al. "Multi-agent learning model with bargaining". *Proceedings of the 2006 winter simulation conference*. IEEE. 2006, pp. 934–940.

[58] J. Rosenski, O. Shamir, and L. Szlak. "Multi-player bandits–a musical chairs approach". *International Conference on Machine Learning*. 2016, pp. 155–163.

[59] S. Shahrampour, A. Rakhlin, and A. Jadbabaie. "Multi-armed bandits in multi-agent networks". *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2017, pp. 2786–2790.

[60] M. Simchowitz and K. Jamieson. Non-Asymptotic Gap-Dependent Regret Bounds for Tabular MDPs. *arXiv preprint arXiv:1905.03814* (2019).

[61] S. Singh, M. Kearns, and Y. Mansour. "Nash Convergence of Gradient Dynamics in General-Sum Games". *In Proceedings of the Sixteenth Conference on Uncertainty in Artificial Intelligence*. Morgan, 2000, pp. 541–548.

[62] J. L. Stimpson and M. A. Goodrich. "Learning to cooperate in a social dilemma: A satisficing approach to bargaining". *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*. 2003, pp. 728–735.

[63] B. Szörényi et al. "Gossip-based distributed stochastic bandit algorithms". *Journal of Machine Learning Research Workshop and Conference Proceedings*. Vol. 2. International Machine Learning Societ. 2013, pp. 1056–1064.

[64] M. S. Talebi and O.-A. Maillard. "Variance-Aware Regret Bounds for Undiscounted Reinforcement Learning in MDPs". *Algorithmic Learning Theory*. 2018, pp. 770–805.

[65] W. Thomson. Nash's bargaining solution and utilitarian choice rules. *Econometrica: Journal of the Econometric Society* (1981), 535–538.

[66] A. Tossou, D. Basu, and C. Dimitrakakis. Near-optimal Regret Bounds for Optimistic Reinforcement Learning using Empirical Bernstein Inequalities. *ERL - ICML'19 Workshop, Submitted to AISTATS* (2020).

[67] A. Tossou, C. Dimitrakakis, and D. Basu. Near-optimal Bayesian Solution For Unknown Discrete Markov Decision Process. *Submitted to STOC* (2020).

[68] A. Tossou et al. A Novel Individually Rational Objective In Multi-Agent Multi-Armed Bandit: Algorithms and Regret Bounds. *Submitted to AAMAS* (2020).

[69] S. Vakili, K. Liu, and Q. Zhao. Deterministic sequencing of exploration and exploitation for multi-armed bandit problems. *IEEE Journal of Selected Topics in Signal Processing* **7**.5 (2013), 759–767.

[70] C.-Y. Wei, Y.-T. Hong, and C.-J. Lu. "Online Reinforcement Learning in Stochastic Games". *Advances in Neural Information Processing Systems 30*. Ed. by I. Guyon et al. 2017, pp. 4987–4997.

[71] Z. Zhang and X. Ji. Regret Minimization for Reinforcement Learning by Evaluating the Optimal Bias Function. *arXiv preprint arXiv:1906.05110* (2019).

[72] M. Zinkevich, A. Greenwald, and M. L. Littman. "Cyclic Equilibria in Markov Games". *Advances in Neural Information Processing Systems 18*. Ed. by Y. Weiss, B. Schölkopf, and J. C. Platt. MIT Press, 2006, pp. 1641–1648. URL: `http://papers.nips.cc/paper/2834-cyclic-equilibria-in-markov-games.pdf`.