Title of Thesis is here

by

Thiruvenkadam Sivaprakasam Radhakrishnan

THESIS

Submitted as partial fulfillment of the requirements
for the degree of Master's in Computer Science
in the Graduate College of the
University of Illinois at Chicago, 2023

Chicago, Illinois

Defense Committee:
Prof. Ian Kash, Chair and Advisor
Prof. Anastasios Sidiropoulos
Prof. Ugo Buy

ACKNOWLEDGMENTS

The thesis has been completed... (INSERT YOUR TEXTS)

YOUR INITIAL

TABLE OF CONTENTS

CHAPTER			PAGE
1	**INTRODUCTION**		1
	1.1	Outline	2
2	**BACKGROUND**		3
	2.1	Game Theory	3
	2.1.1	Problem Domains and Representations	3
	2.1.2	Solution Concepts	3
	2.2	Reinforcement Learning	4
	2.2.1	Policy gradient methods	4
	2.2.1.1	Softmax Policy Gradients	4
	2.2.2	PPO	4
	2.3	Online Learning	4
	2.3.1	FoReL	4
	2.3.2	Hedge	4
3	**MAGNETIC MIRROR DESCENT**		5
	3.1	Mirror Descent	5
	3.1.1	MDPO	5
	3.2	MMD	5
	3.2.1	Zero-sum games and Variational Inequalities	6
4	**FASTER MMD AND MDPO**		9
	4.1	Faster-MMD	9
	4.2	Neural Replicator Dynamics (NeuRD)	9
	4.2.1	FMMD-N	10
	4.3	Extragradient updates	10
	4.3.1	FMMD-EG	10
	4.4	Optimism	10
	4.4.1	Optimistic Mirror Descent	10
	4.4.2	OFMMD	10
5	**EXPERIMENTS**		11
	5.1	Experimental Domains	11
	5.2	Evaluation Methods	11
	5.3	Tabular MMD Experiments	11
	5.3.1	Results	11
	5.4	Neural MMD Experiments	11

TABLE OF CONTENTS (Continued)

CHAPTER **PAGE**

5.4.1 Results 11

APPENDICES 12
Appendix A 13
Appendix B 14

CITED LITERATURE 15

LIST OF TABLES

TABLE **PAGE**

LIST OF FIGURES

FIGURE **PAGE**

LIST OF NOTATIONS

θ	Parameters of a function approximator.

SUMMARY

Put your summary of thesis here.

CHAPTER 1

INTRODUCTION

In this work, we study two mirror-descent based reinforcement learning algorithms and propose novel improvements to them.

1.1 Outline

The rest of the thesis is organized as follows. We begin by providing some background and definitions in section 2 that are useful for the understanding of the algorithms and methods described in section 3 and 4. Section 3 introduces Mirror Decsent and expands on Mirror Descent based methods for solving Reinforcement learning problems. Section 4 discusses combining novel improvements on top of these methods and discusses their structure and the expected effects. In Section 5, we dive into some experimental results and discuss the performance of these algorithms in different settings. We then close the thesis with some discussion.

CHAPTER 2

BACKGROUND

We begin by providing some necessary background in Game Theory, Online Learning, and Reinforcement learning to make the reader familiar with the concepts required to follow the ideas discussed in the following sections.

2.1 Game Theory

- what is game theory? - why is it useful? - how is it related?

2.1.1 Problem Domains and Representations

- how are problems typically represented in game theory? - what are the usual representations? Normal Form games, Extensive Form games, Sequence Form. - how is this relevant?

2.1.2 Solution Concepts

- what are solution concepts? - what are the common solution concepts? Nash equilibrium, Quantal response equilibrium. - what are the relevant information related to solution conepts for this work? Existence of a nash equilibrium Uniqueness of QRE

2.2 Reinforcement Learning

2.2.1 Policy gradient methods

2.2.1.1 Softmax Policy Gradients

2.2.2 PPO

2.3 Online Learning

- what is online learning?

Online learning is the study of designing algorithms that use historical knowledge in predicting actions for future rounds while trying to minimize some loss function in an adaptive (possibly adversarial) setting.

- why is it useful? - why is it relevant here?

2.3.1 FoReL

- what is forel? - relevant info?

2.3.2 Hedge

- what is hedge?

CHAPTER 3

MAGNETIC MIRROR DESCENT

3.1 Mirror Descent

In this section we introduce Mirror Descent, and discuss MDPO, and MMD - two mirror descent-based reinforcement learning algorithms.

A disadvantage of FoReL 2.3.1 in solving online learning problems is that, there is a minimization at every step. Mirror descent overcomes this disadvantage by using a recursive update rule that does not required us to perform a minimization at every step.

3.1.1 Mirror Descent Policy Optimization

3.2 Magnetic Mirror Descent

Magnetic mirror descent (MMD) [1] is a last-iterate equilibrium approximation algorithm for two-player zero-sum games that is an extension of mirror descent with entropy regularization.

The idea behind MMD begins with the observation that solving for QRE in two-player zero-sum games can be reformulated as the solution to a negative entropy regularized saddle point problem as follows,

$$\min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} \alpha g_1(x) + f(x,y) + \alpha g_2(y), \tag{3.1}$$

where $\mathcal{X} \subset \mathbb{R}^n$, $\mathcal{Y} \subset \mathbb{R}^m$ are closed and convex, and $g_1 : \mathbb{R}^n \to \mathbb{R}$, $g_2 : \mathbb{R}^m \to \mathbb{R}$, $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$.

The solution (x_*, y_*) to Equation 3.1, is a Nash equilibrium in the reguarlized game with the following best response conditions,

$$x_* \in \arg\min_{x \in \mathcal{X}} \alpha g_1(x) + f(x, y_*) \tag{3.2}$$

$$y_* \in \arg\min_{y \in \mathcal{Y}} \alpha g_2(y) + f(x_*, y) \tag{3.3}$$

3.2.1 Zero-sum games and Variational Inequalities

MMD reframes the solution to QRE as a variational inequality problem.

Definition 1 *Given $\mathcal{Z} \subseteq \mathbb{R}^n$ and mapping $G : \mathcal{Z} \to \mathbb{R}^n$, the variational inequality problem VI $(\mathcal{Z}, G)$ is to find $z_* \in \mathcal{Z}$ such that,*

$$\langle G(z_*, z - z_*) \rangle \geq 0 \quad \forall z \in \mathcal{Z}.$$

The optimality conditions are equivalent to VI$(\mathcal{Z}, G)$, where $G = F + \alpha \nabla g$, $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, and $g : \mathcal{Z} \to \mathbb{R}$.

Now, the solution the the VI problem ($z_* = (x_*, y_*)$), corresponds to the solution of the saddle point problem stated in Equation 3.1, and satisfies the best response conditions Equation 3.2 and Equation 3.3.

The algorithm that the authors propose is a non-Euclidean proximal gradient method to solve the VI $(\mathcal{Z}, F + \alpha \nabla g)$ problem.

We now restate the main algorithm as stated in Sokota et. al, [1] (3.1),

---

AlgorithmMMD Starting with $z_1 \in \text{int dom } \psi \cap \mathcal{Z}$, at each iteration t do

$z_{t+1} = \arg\min_{z \in \mathcal{Z}} \eta(\langle F(z_t), z\rangle + \alpha g(z)) + B_\psi(z, z_t)$.

With the following assumptions, z_{t+1} is well defined:

- ψ is 1-strongly convex with respect to $\|.\|$ over $\mathcal{Z}$, and for any l, stepsize $\eta > 0$, $\alpha > 0$, $z_{t+1} = \arg\min_{z \in \mathcal{Z}} \eta(\langle l, z\rangle + \alpha g(z)) + B_\psi(z; z_t) \in \text{int dom } \psi$.
- F is monotone and L-smooth with respect to $\|.\|$ and g is 1-strongly convex relative to ψ over $\mathcal{Z}$ with g differentiable over int dom ψ.

Algorithm 1 provides the following convergence guarantees,

Assuming that the solution z_* to the problem VI $(\mathcal{Z}, F + \alpha\nabla g)$ lies in the int dom ψ, then the algorithm 1 guarantees

$$B_\psi(z_*; z_{t+1}) \leq \left(\frac{1}{1+\eta\alpha}\right)^t B_\psi(z_*; z_1),$$

if $\alpha > 0$, and $\eta \leq \frac{\alpha}{L^2}$.

In single-agent settings MMD's performance is competitive with PPO in Atari and MuJoCo environments. And, in the multi-agent setting the performance of tabular MMD is on par with CFR, but worse than CFR+.

CHAPTER 4

FASTER MMD AND MDPO

4.1 Faster-MMD

We now propose a few modified versions of Magnetic Mirror Descent, and MDPO

4.2 Neural Replicator Dynamics (NeuRD)

Neural Replicator Dynamics (NeuRD) [2] is a model-free sample-based algorithm that applies function approximation to Replicator Dynamics. Replicator Dynamics is an idea from Evolutionary game theory (EGT) that defines operators to update the dynamics of a population in ord to maximize some pay-off defined by a fitness function.

The single-population replicator dynamics is defined by the following system of differntial equations:

$$\dot{\pi}(a) = \pi(a)[u(a, \pi) - \bar{u}(\pi)], \forall a \in \mathcal{A} \tag{4.1}$$

Hennes et. al, [2] show equivalence between Softmax Policy gradients 2.2.1.1 and continuous-time Replicator Dynamics [2, THEOREM 1, on p5].

NeuRD can be applied to reinforcement learning as a single line change to the Softmax Policy gradient. This can be seen as applying a fix to the update of the Softmax Policy gradient algorithm to make it more responsive to changes in a non-stationary environment. We refer to this as the "NeuRD-fix".

4.2.1 FMMD-N

The first modified version of MMD, and MDPO is obtained applying the NeuRD-fix to these algorithms. Below, we derive the gradient of the behavioral form update rule for these algorithms, and show where the fix applies in these updates.

4.3 Extragradient updates

The Extragradient method was first introduced by G.M.Korpelevich [3]. EG is a classical method for solving smooth and strongly convex-concave bilinear saddle point problems with a linear rate of convergence. Extragradient and Optimistic Gradient Descent Ascent methods have been shown to be approximations of proximal-point method for solving saddle point methods [4].

4.3.1 FMMD-EG

In this section, we outline a version of MMD with extragradient updates that we call *FMMD-EG*.

4.4 Optimism

4.4.1 Optimistic Mirror Descent

4.4.2 OFMMD

CHAPTER 5

EXPERIMENTS

5.1 Experimental Domains

Tabular Experiments: Normal Form games: Perturbed RPS

Neural Experiments: Kuhn Poker Abrupt Dark Hex (3x3) Phantom Tic-tac-toe

5.2 Evaluation Methods

Exact exploitability Apporximate exploitability

5.3 Tabular MMD Experiments

5.3.1 Results

5.4 Neural MMD Experiments

Outline:

- Implementation details (RLLib, PPO modifications, GAE)
- Neural network architecture, hyperparameters

5.4.1 Results

APPENDICES

Appendix A

SOME ANCILLARY STUFF

Ancillary material should be put in appendices.

Appendix B

SOME MORE ANCILLARY STUFF

[5]

CITED LITERATURE

1. Sokota, S., D'Orazio, R., Kolter, J. Z., Loizou, N., Lanctot, M., Mitliagkas, I., Brown, N., and Kroer, C.: A Unified Approach to Reinforcement Learning, Quantal Response Equilibria, and Two-Player Zero-Sum Games. In *The Eleventh International Conference on Learning Representations* , February 2023.

2. Hennes, D., Morrill, D., Omidshafiei, S., Munos, R., Perolat, J., Lanctot, M., Gruslys, A., Lespiau, J.-B., Parmas, P., Duenez-Guzman, E., and Tuyls, K.: Neural Replicator Dynamics, February 2020.

3. Korpelevich, G. M.: The extragradient method for finding saddle points and other problems. *Matecon* , 12:747–756, 1976.

4. Mokhtari, A., Ozdaglar, A., and Pattathil, S.: A Unified Analysis of Extra-gradient and Optimistic Gradient Methods for Saddle Point Problems: Proximal Point Approach. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics* , pages 1497–1507. PMLR, June 2020.

5. Farine, D. R., Strandburg-Peshkin, A., Couzin, I. D., Berger-Wolf, T. Y., and Crofoot, M. C.: Individual variation in local interaction rules can explain emergent patterns of spatial organization in wild baboons. *Proceedings of the Royal Society of London B: Biological Sciences* , 284(1853), 2017.