PURDUE UNIVERSITY

Reinforcement Learning based efficient spectrum sensing and access policies for Cognitive Radio Networks

by

Bharath Keshavamurthy

A thesis proposal submitted in partial fulfillment for the
Master of Science degree

under the guidance of
Professor Nicolo Michelusi
School of Electrical and Computer Engineering

September 2018
v1.0.0

Declaration of Authorship

I, BHARATH KESHAVAMURTHY, declare that this graduate thesis proposal titled, 'REINFORCEMENT LEARNING BASED EFFICIENT SPECTRUM SENSING AND ACCESS POLICIES FOR COGNITIVE RADIO NETWORKS' and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a Master of Science degree at the School of Electrical and Computer Engineering, Purdue University.
- Where any part of this thesis proposal has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis proposal is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: *Bharath Keshavamurthy*

Date: 23-September-2018

"What I cannot create, I do not understand."

Richard Phillips Feynman

PURDUE UNIVERSITY

Abstract

Master of Science

by Bharath Keshavamurthy

This graduate thesis proposal document details the numerous open problems in the design and operation of Cognitive Radio Networks. Additionally, the document also outlines solutions and solution methodologies to solve these open problems. There is a need for an over-arching spectrum sensing and scheduling framework that operates in both collaborative and non-collaborative radio environments, with or without the presence of Fusion Centres. The Secondary User (SU) radio environment should be able to discern the situation it is in at any given point in time and determine the best possible action policy in order to maximize the throughput of this SU network while keeping PU interference to a minimum. Reinforcement Learning techniques coupled with Supervised and Unsupervised learning algorithms can be employed to tackle this requirement as these "intelligent" frameworks, also known as "expert systems", most effectively model the dynamism of the Cognitive Radio ecosystem. For instance, Clustering algorithms can be employed to facilitate neighbour discovery, trust-based heuristics can be incorporated to 'elect' cluster-heads for distributed sensing in ad-hoc SU networks, SVM-based supervised learning techniques can be used to determine PU channel access schemes, autonomous participation strategies can be integrated at secondary radio nodes based on cost-reward Bayesian game heuristics, and Bandits, Markov Decision Processes (MDPs), and Model-Free Learning agents can be leveraged to solve for optimal action policies. Furthermore, the secondary considerations of reducing control channel overhead, reducing the dedicated spectrum sensing time per SU, and improving the energy efficiency of radio nodes can be consolidated into the overall optimization problem. The purpose of this research proposal is to carry out rigorous functional and performance evaluations of the proposed framework against existing state-of-the-art and conclusively prove that incorporating an adaptive, hierarchical architecture leveraging intelligent sensing and scheduling policies based on Reinforcement Learning and Supervised/Unsupervised Learning techniques, out-performs any existing state-of-the-art. Finally, the proposal also details intentions of prototyping and testing certain aspects of the framework in emulated radio environments with other "unknown" collaborative/competing radio nodes, such as Scenario-based Channel Emulation on the DARPA SC2 Colosseum.

Contents

List of Figures

List of Tables

Dedicated to John Forbes Nash Jr. ...

Chapter 1

Introduction

The problem of spectrum scarcity has been in the spotlight for the past few years primarily owing to the advent of fifth-generation wireless technologies and the increased penetration of the "Internet of Things" into our day-to-day lives. Both these advances rely on Massive Machine to Machine communication which would, if they're fully prevalent, impose enormous strain on the available spectrum. The amount of spectrum available for commercial use is limited and Communication Commissions around the world such as the Federal Communications Commission (FCC) typically license portions of this spectrum to operators for huge sums of money. The enormous additional burden imposed by next-generation wireless technologies and the Internet of Things cannot be accommodated by conventional licensing strategies. This problem has brought up a need for dynamic spectrum access policies, thereby proliferating research in Cognitive Radio Networks. Dynamic Spectrum Access and Management strategies in Cognitive Radio Networks facilitate concurrent utilization of the spectrum while keeping interference with the primary, licensed users to a minimum. The proposed research aims to deliver an adaptive, hierarchical, intelligent spectrum sensing and spectrum access policy-driven framework for Cognitive Radio Networks leveraging techniques from Bandits, Reinforcement Learning, Supervised/Unsupervised Learning, and Game Theory. In general, the optimization problem can be modelled as maximizing the overall throughput of the Secondary User network while keeping the missed detection probability below a certain threshold. Moreover, secondary objectives of the heuristics involved in the proposed framework include energy efficiency of the SUs, minimizing control channel overhead, and limiting false alarm rate. The term "Missed detection", as the name suggests, refers to the system incorrectly identifying an occupied sub-band as being idle while the term "False alarm rate" refers to the system incorrectly identifying an idle sub-band as being occupied.

This graduate research proposal is broken down into five phases: RL-agents in Centralized, Collaborative CRNs (inclusion in the BAM! Wireless collaboration framework employing information extracted from CIL client-server messages and peer-to-peer messages); Distributed learning heuristics in decentralized ad-hoc collaborative networks along with optimal neighbor discovery; Opportunistic spectrum access in competing radio resource utilization environments; Functional and Performance comparison of algorithms that employ CNNs/SVMs for PU behavior modelling and SU scheduling, algorithms that model the spectrum access and autonomous participation behavior in spectrum sensing using game-theoretic heuristics, and Reinforcement Learning based multi-agent, multi-band spectrum sharing policies; and Coalescence of various approaches into a best-policy framework along with possible applications in the 5G and IoT landscape while also leveraging the programmability and controllability facets of Software Defined Networks to deliver a complete, adaptive, hierarchical, and intelligent framework.

Under-utilized licensed spectrum is a time-frequency-location varying resource influenced by PU behavior, radio wave propagation, and signal attenuation. Dynamic Spectrum Access (DSA) policies in Cognitive Radio Networks (CRNs) are concerned with identifying the temporal and spatial holes in the spectrum and capitalizing on them while ensuring non-interference with the PUs. Dynamic Spectrum access can be broadly categorized into Collaborative Spectrum Access and Opportunistic Spectrum Access.

Collaborative or Cooperative spectrum sensing plays a crucial role in identifying the spectrum holes by mitigating the effects of fading and shadowing. Collaborative Spectrum Sensing can be achieved in three ways: Using a centralized aggregation approach employing Fusion Centres, A semi-centralized clustered topology where in highly correlated SUs are clustered and one node among them is "nominated" as the cluster head which operates on the next tier as just another radio node in the distributed radio environment, or A completely distributed ad-hoc Cognitive Radio environment in which SUs make local decisions to sense and access portions of the spectrum based on their observations and received test statistics from their neighbours. Broadly speaking, the optimization problem involves maximizing the SU network throughput while keeping the missed detection probability under a certain threshold.

Non-Collaborative Cognitive Radio Networks deal with environments in which all the SUs opportunistically try to access the spectrum and complete their network flows. "Opportunistic access" among competing radio nodes brings in a completely new set of problems to consider such as mandated acknowledgements, intelligent back-off policies, modelling the effects of fading and shadowing into observations of PU activity, and distributed learning policies without any prior knowledge of the radio environment or without any kind of information exchange among the radio nodes.

Chapter 2

Literature Survey

Although there have been a few research publications detailing policies and strategies for Dynamic Spectrum Access in Cognitive Radio Networks, there are numerous open problems that are yet to be solved. For instance, clustering strategies to devise groups of highly correlated SUs for collaboration is one approach to distributed spectrum sensing which has not been attempted in existing state of the art. Furthermore, a large number of research publications in the area of Cognitive Radio Networks assume some kind of Channel State Information or prior knowledge which may not be the case in completely ad-hoc radio environments. Additionally, the entire state-of-the-art assume independence of frequency sub-bands with respect to PU access- this is not always the case and this one huge open problem which needs to be solved by devising learning strategies which produce optimal or nearly-optimal outcomes. The end goal of the proposed research endeavour is to produce an optimal, well-defined, intelligent, adaptive framework that can be deployed in next-generation wireless networks irrespective of the underlying network topology and traffic scenarios. Some of the open problems observed in the existing state-of-the-art are listed in Section 2.1. This research proposal aims to solve these open problems and coalesce the solutions into an intelligent, over-arching framework that adapts itself to changing network topologies and radio environment conditions. Section 2.2 lays down brief analyses of existing literature in this domain.

2.1 Open Problems

- Optimal neighbour discovery strategies in semi-centralized or distributed Cognitive Radio Networks to facilitate collaboration and thereby mitigate the effects of fading and shadowing that may creep into observations of PU activity.

- The evaluation of frameworks which involve compliant communications among peers over a dedicated collaboration network which facilitate a better picture of spectrum utilization in the Cognitive Radio Network.

- Analysis of the potential impact on control channel overhead caused by the dissemination of test statistics among radio nodes in distributed collaborative environments or to and from the Aggregator in centralized collaborative environments. The optimal policy should conclusively reduce the control channel overhead compared to policies in the existing state-of-the art.

- Tuning control knobs such as Diversity Order, Missed Detection Probability, False-Alarm Rate, Fusion rules, Number of sub-bands to be sensed per node, etc. and observe corresponding variations in control channel overhead, energy efficiency of radio nodes, aggregation latency, and SU network throughput.

- How do we model the radio environment without the assumption of independence between sub-bands conditioned on the number of PUs in the network?

- How do we factor in the impact of fading and shadowing on local PU observations when SUs do not exchange any information among them in non-collaborative Cognitive Radio Networks? Also, there is a need for intelligent back-off strategies and operational scheduling policies in these radio environments where each SU is trying to complete its network flows without interacting with other SUs in the network.

- Estimation of state transition probabilities when modelling the spectrum sensing policy optimization using Markov Decision Processes (MDPs).

- There is a need for an intelligent, adaptive, hierarchical framework that learns the radio network behaviour irrespective of the topology, application, and traffic scenarios, and produces an optimal or nearly-optimal policy which maximizes the throughput of the SU network while limiting PU interference along with constraints on control channel overhead, decision latency, and energy efficiency of SUs. A coarse approach to devise this adaptive framework would be to leverage the programmability of Software Defined Networking (SDN) and use heuristics in the Application layer to modify radio node operational parameters in the Data plane through the Control plane using simple protocols such as REST, SOAP, and CLI. A finer approach would be to embed this adaptive intelligence within the Cognitive Radio Network.

2.2 State-of-the-Art

- Reference 8 details optimal neighbour discovery heuristics using finite Markov Decision Processes in Cognitive Radio Ad-Hoc Networks (CRAHNs). The initiating SU requests it's one-hop neighbours for local test statistics and aggregates them to get a spectrum utilization map in the "protected region". The PU activity is modelled as a two-state Birth-Death process in which the transitions are Poisson processes. The observations from this sensing operation are disseminated to all the neighbours. This cooperative sensing strategy is iterated over N episodes to figure out an optimal policy using Reinforcement Learning.

- Reference 1 details the use of an ϵ-greedy algorithm to devise a Reinforcement Learning based spectrum sensing policy optimization. Here, the optimal sensing assignments are learned using Q-value optimization along with a solution for the exploration-exploitation trade-off using the ϵ-greedy algorithm. Pseudo-Random sequences called Frequency Hopping codes are employed to facilitate exploration.

- Reference 4 outlines the use of a Multi-agent Multi-band distributed Reinforcement Learning approach using SARSA with Linear Function Approximation for Dimensionality Reduction.

- Reference 9 details the use of Multi-Armed Bandit frameworks to enable reliable and efficient spectrum access in Opportunistic Cognitive Radio Networks. Here, g-statistic values are computed for each orthogonal channel in the spectrum of interest and the SU selects a channel with the highest g-statistic. However, this approach leads to large number of collisions among SUs which have decided to access the same channel. Hence, intelligent back-off strategies and mandated acknowledgements are essential here.

- Reference 17 outlines the use of Support Vector Machines (SVMs) to differentiate between various PU channel access schemes such as TDMA, Aloha, Slotted Aloha, and CSMA/CA. Reference 10 details the use of fourth-order cumulant-based classifiers to distinguish between TDMA, OFDMA, and CDMA. Reference 10 also employs a cumulant sample variance based collision detector to detect contention-based channel access schemes. Other research endeavours such as the one described in 18 uses Supervised Learning algorithms to categorize PU activity into subsets of the 802.11 standard.

- Reference 12 describes the use of various heuristics within three proposed SU scheduling strategies- Sequential, Parallel, and Sequential-Parallel. Reference 11 proposes the use of soft and hard reports algorithms to learn the footprints of the

PUs. Detection of active components is done by comparing the received energy observations with the learned candidate and confirmed components. The paper also lays down heuristics for confirming, deleting, and merging source components corresponding to PUs in the network.

Chapter 3

Centralized Collaborative Cognitive Radio Environments

3.1 Overview

The first phase of the research proposal detailed in this document deals with the incorporation of Bandit frameworks and Reinforcement Learning techniques to devise optimal spectrum sensing and access policies in Centralized Collaborative Cognitive Radio environments. Additionally, novel fusion heuristics are proposed to incorporate the CIL client-server messages and peer-to-peer messages into the global decision-making at the Fusion Centre, also known as the Aggregator node or the Gateway node. The end goal of this phase is to conclusively prove that the proposed framework out-performs existing state-of-the-art by evaluating its functionalities and performance metrics in emulated radio environments such as the various RF/Traffic scenarios on the DARPA SC2 Colosseum.

The proposed research aims to incorporate the use of multi-band, multi-user, centralized, collaborative, Reinforcement Learning based spectrum sensing policy for efficient spectrum sharing in the DARPA SC2 radio environment. The proposed policy implicitly learns the Primary User (PU) behavior over time by directly interacting with the radio environment, i.e. learning the best action policy based on Q-value optimization. The optimization problem is to maximize the throughput of the Secondary User (SU) network while keeping the missed detection probability below a certain threshold. Secondary objectives of the heuristics involved in the system include energy efficiency of the SUs, minimizing control channel overhead, and limiting false alarm rate.

Under-utilized licensed spectrum is a time-frequency-location varying resource influenced by PU behavior, radio wave propagation, and signal attenuation. Dynamic Spectrum Access (DSA) policies in Cognitive Radio Networks (CRNs) are concerned with identifying the temporal and spatial holes in the spectrum and capitalizing on them while ensuring non-interference with the PUs. Collaborative or Cooperative spectrum sensing plays a crucial role in identifying these spectrum holes by mitigating the effects of fading and shadowing. In a collaborative spectrum sensing environment, multiple SUs sense the same sub-bands and send their local statistics to an Aggregator node which may be just another standard radio node or a dedicated Fusion Center. The Aggregator, upon receiving the test statistics for the most optimal sub-bands from groups of spatially and temporally optimal Secondary Users along with the Report/Violation/Spectrum-Utilization messages received over the DARPA SC2 collaboration channel according to the CIL specifications, constructs a global view of the spectrum at that time step. To reiterate, the optimization problem involves maximizing the SU network throughput while keeping the missed detection probability under a certain threshold.

3.2 Application and Emulation in the DARPA SC2 Radio environment

In competition or scrimmage events, only one SRN (LXC container) can be designated a Gateway Node. The container should use the presence of the col0 interface to determine that they are the gateway node. This proposal intends to call the gateway SRN the Aggregator while the non-gateway SRNs are simply termed SUs. The Aggregator is a part of the SC2 collaboration network consisting of a Collaboration server which serves as a PUB-SUB framework indicating peer-entry, peer-removal, and other specific collaboration messages. The SC2 collaboration network is a /24 broadcast domain over a link-local routes only wired IP network with the SC2 collaboration server at the center of it broadcasting CIRN Interaction Language (CIL)-specific collaboration messages to the gateway SRNs in the network.

3.3 Anticipated Contributions

The incorporation of Reinforcement Learning algorithms (RL) into the spectrum sensing framework allows the system to learn the radio environment without the need for prior dynamic modelling. The problem of modelling an extremely dynamic radio environment is reduced down to an optimization problem of maximizing SU network throughput while having a constraint on the missed detection probability. The exploration-exploitation

heuristics modelled into RL algorithms makes them very suitable for reaching optimal action policies when we do not have prior information on the reward distributions of the environment were modelling. Intuitively, the optimization problem can be stated as finding the most optimal sub-bands and the most optimal (SU, sub-band) sensing assignments in order to ensure all the SU flows are successfully fulfilled without any need for re-transmissions and without violating the SC2 PU interference constraints. The dissemination of binary decisions/statistics from the SUs to the Aggregator greatly reduces the control channel overhead as opposed to sharing the entire Channel State Information (CSI). The incorporation of the Diversity Order metric (D) opens up the controllability of the framework. Even the use of simple ϵ-greedy algorithms to find near-optimal sensing assignments provide impressive results as seen in 1. Varying complexities of fusion rules can be incorporated at the Aggregator node and this provides us with a incredible amount of control & management flexibility. The use of pseudo-random frequency hopping codes as discussed in 2 constitutes the exploration aspect of the sensing policy which prevents the policy improvement algorithm from settling on sub-optimal actions. This framework serves as a logical starting point to tackle more complicated problems in this arena such as distributed sensing in ad-hoc networks, mobility of nodes, neighbor discovery, opportunistic spectrum access in competing radio networks, etc.

Phase 1 involves three approaches to solve the optimization problem for Collaborative Cognitive Radio Networks in a Centralized topology. Each of these approaches are explained in detail below.

3.4 Bandit frameworks at the Aggregator

3.4.1 Advantages

- No complex modelling of state and action spaces are needed
- No prior knowledge about the system is assumed

3.4.2 Problem Formulation

The problem of frequency band selection for sensing can be modelled as a Restless Multi-Armed Bandit one such that the decision-maker (the bandit) chooses L out of N frequency bands to sense (L out of N arms to pull) where $L \geq 1$ and $L \leq N$. Restless Multi-Armed Bandit (R-MAB) frameworks model the DSA scenario in CRNs accurately because the states of the frequency bands which are not sensed in a particular time-step change as opposed to a stationary MAB formulation where in the reward distributions

of un-played arms remain the same. Moreover, the lack of prior information about the reward distributions of various frequency bands brings in the known problem of exploration v/s exploitation which can be solved very well using numerous Multi-Armed Bandit algorithms such as UCB, Posterior Sampling, ϵ-greedy, and contextual bandit algorithms such as LinUCB. UCB (employs exploration bonus term in optimal policy selection) and Posterior Sampling (requires some prior knowledge about the reward distribution to update the posterior reward distribution and choose the optimal action policy) have asymptotic logarithmic regret which intuitively means that these algorithms strike the perfect balance between exploration and exploitation. LinUCB is a contextual bandit framework that models the current state information into optimal policy selection by developing a state-arm embedding and then using ridge regression to estimate the theta matrix which is then included in the policy selection algorithm along with an exploration bonus.

3.4.3 System Model

The spectrum of interest is assumed to be divided into NB sub-bands of identical or different bandwidths. Any of these bands can be occupied by any PUs at any time. The problem is much more difficult and dynamic when the assumption of well-demarcated sub-bands is removed (well tackle this at a later stage). Based on the physical capabilities of the SUs, one SU can sense up to Ks bands simultaneously. Let the number of cooperating SUs be N_s. The SU operation is divided into Sensing timeslots and Transmission time-slots. In each sensing time slot, the SUs sense the bands assigned to them by the Aggregator and send their binary decisions about these bands to the Aggregator along with their transmission requirements such as flow_enabled flag, estimated_flow_throughput, SRN_ID, and other relevant meta-data. The Aggregator, upon receiving the SU decisions and the CIL messages over the collaboration channel employs simple fusion rules such as the OR rule or the K-out-of-N rule or more complex fusion rules which can be captured in decision-processing frameworks such as Apache NiFi, Apache JEXL, Drools, etc to make more-informed, global decisions about the state of the sensed frequency bands.

The Aggregator employs two Q-value optimization algorithms: one for the frequency bands and the other for the (SU, Freq. band) assignment combinations. The Q-value optimization algorithm for each frequency band 'b' in the spectrum of interest has the following reward assignment:

$$r_{k+1}(b) = Throughput\ of\ b,\ if\ the\ decision\ at\ the\ Aggregator\ finds\ it\ to\ be\ free$$

$$r_{k+1}(b) \;=\; 0,\; if\; the\; decision\; at\; the\; Aggregator\; finds\; the\; band\; to\; be\; occupied$$

The Q-values for the (SU, Freq. band) assignment pairs are optimized based on the following reward heuristics:

$$r_{k+1}(s,b) \;=\; the\; local\; binary\; decision,\; if\; the\; decision\; at\; the\; Aggregator\; for\; b\; is\; 1$$

$$r_{k+1}(s,b) \;=\; Q_k(s,b),\; if\; the\; decision\; at\; the\; Aggregator\; for\; band\; b\; is\; 0$$

These reward heuristics are picked from 1 and they need to be changed based on the convergence/near-convergence of our algorithms in the DARPA SC2 radio environment. After making a decision on the availability of a certain band, the Aggregator RL agent updates the Q-function for that particular band based on the following equation:

$$Q_{k+1}(a) \;=\; Q_k(a) \;+\; \alpha[r_{k+1}(a)Q_k(a)]$$

where, α = a constant step-size factor such that,

$$0 \;\leq\; \alpha \;\leq\; 1$$

As α is made larger, more emphasis is placed on recent rewards while as α approaches 0, the algorithm will push emphasis on rewards obtained in the past. So, this parameter will be an important control knob in the system. Additionally, the Aggregator also updates the Q-values of the (SU, Freq. band) sensing assignment pairs using the above equation and reward heuristics outlined in the previous slide. In a simple ϵ-greedy MAB algorithm, with a probability (1 - ϵ) of the system can be made to choose L out of N_s frequency bands to sense based on the throughput requirements, i.e. the flows assigned to the BAM! Wireless network nodes, by simply selecting the L bands with the highest Q-values and their correspondingly optimal sensing assignments. With a probability of ϵ, the frequency-band selection and sensing assignments are done either uniformly at random or by employing pseudo-random frequency hopping codes as outlined in 2.

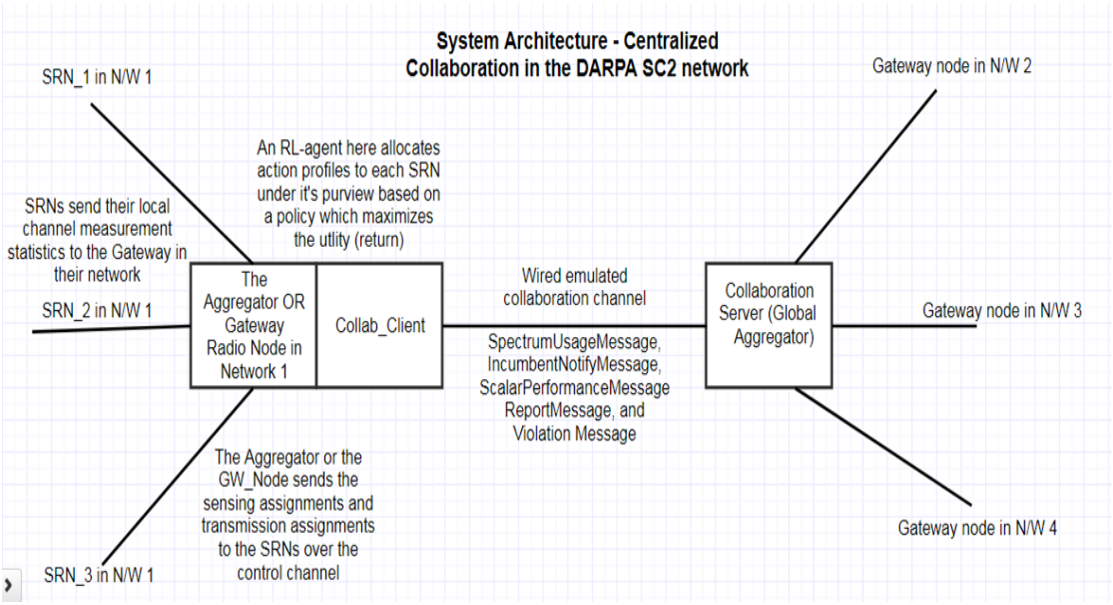

FIGURE 3.1: System Architecture

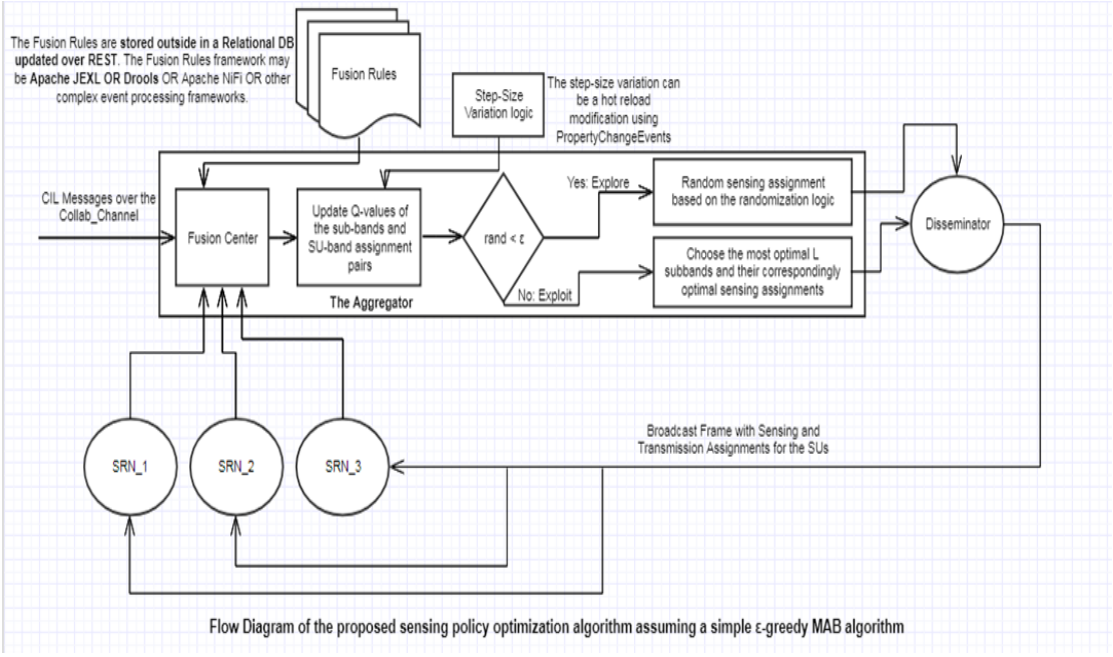

FIGURE 3.2: Sensing Policy Optimization Flow Diagram

Figure 3.1 shows the architecture of the Restless Multi-Armed Bandits framework based Centralized Collaborative Cognitive Radio environment.

Figure 3.2 shows the flow diagram of the sensing policy optimization approach using a Restless Multi-Armed Bandits framework in Centralized Collaborative Cognitive Radio Networks.

3.4.4 Optimization Problem

The bandit algorithm at the Aggregator is going to choose L sub-bands out of N_s possible sub-bands to be sensed at that time step by selecting those sub-bands which possess the highest Q-values (provide maximum utility to the system). After choosing these $L < N_B$ sub-bands, the exploration-exploitation engine kicks in which chooses with a probability of *epsilon*, a random sensing assignment based on the defined randomization logic (can be as simple as a random selection of frequency hopping codes with fixed or varying diversity orders) AND, with a probability of $(1 - \epsilon)$, chooses a sensing assignment according to the following optimization problem:

$$\min_{X} \sum_{b \in B} \sum_{s \in \mathcal{S}} w_s x_{sb} \; such \; that,$$

$$\hat{P}^{b}_{miss,Global}(X) \leq P^{b}_{miss,predefined} \; and,$$

$$\sum_{b \in B} x_{sb} \; \leq \; K_s \; where, \; x_{sb} \in \{0, 1\}$$

Here, w_s is the weight assigned to each SU $s \in \mathcal{S}$, X is N_s X L sensing assignment matrix x_{sb} is an element of X which is set to 1 if the SU $s \in \mathcal{S}$ is assigned to sense $b \in B$, else it is set to 0.

Reference 1 uses hard-decision combining of multiple Neyman-Pearson detectors to solve the optimization problem outlined in the previous slide, i.e. maximize detection probability. False alarm rate constraints are included in Neyman-Pearson detector heuristics. Using simple OR-Fusion Rules, the optimization problem can be converted into a linear BIP problem as follows which is solved in 1 using Branch-and-Bound searches and an Iterative Hungarian algorithm.

$$\min_{x} \; w^T x, \; such \; that \; Ax \leq c,$$

where, A is the constraint matrix c is a constraint vector, and x is a binary vector of X.

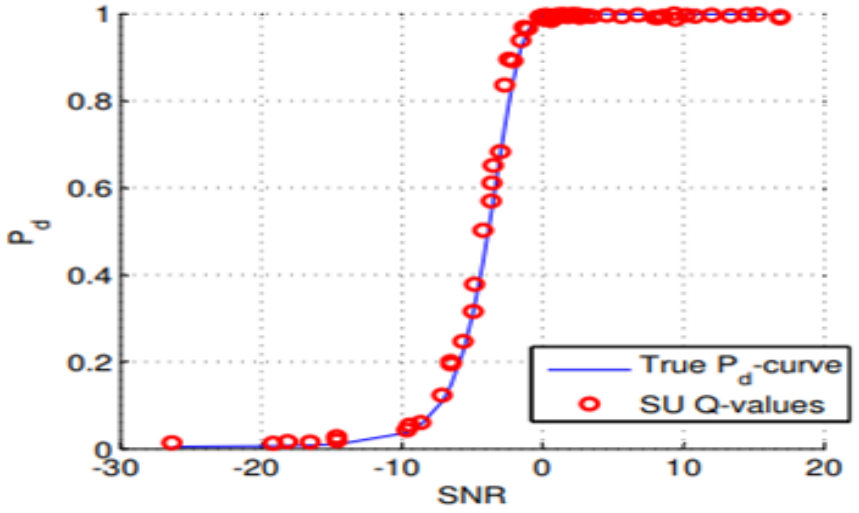

FIGURE 3.3: Sensing Policy Convergence with the true probability detection curve

The analytical expressions in 1 and 2 show that as the number of interactions with the radio environment approach $+\infty$, the optimization problem outlined in the previous slide converges in the same manner as the Q-value convergence for the (SU, Freq. band) pair assignment. The following figure from 1 shows that the Q-values align with the true probabilities of detection.

3.5 Markov Decision Process based RL agent at the Aggregator

3.5.1 Key Differentiators

- State space and Action space definitions
- Dimensionality reduction (if needed)
- Solve the optimization problem using Dynamic Programming, i.e. Policy Iteration (Policy Evaluation + Policy Improvement)
- Estimate the state transition probabilities using stochastic approximation theory

3.5.2 Overview

An MDP-based RL agent running on the Aggregator node chooses optimal sensing assignments for the N_s SUs with respect to the NB frequency bands in the spectrum of interest. Local test statistics from the SUs along with the statistics from the CIL messages are fused into a global spectrum state map at the Aggregator node. As mentioned in the previous approach, fusion rules are made to sit outside the framework for easy hot-reloading of system behavior (leverage the PropertyChangeEvent design pattern in software development practices). The optimal policy at the Aggregator involves choosing an action at each state which maximizes the system utility, i.e. the throughput of the secondary network.In scenarios where the state space turns out to be too large to converge, dimensionality reduction techniques based on state space approximation can be employed to ensure optimal or near-optimal policy selection. Dynamic programming techniques are employed to solve the optimization problem: Policy Evaluation and Policy Improvement.

3.5.3 The MDP State Space

The state space is a set of all possible binary code-words of size N_B. The state space represents the sub-band specific spectrum occupancy map of the spectrum under observation. Each sub-band in the spectrum of interest is represented by a sub-state variable ω, which can take binary values, i.e. $\omega \in \{0, 1\}$.

$$\mathcal{S} \;=\; \{\; \{\omega_i\} :\; \omega \in \{0,1\} \; and \; i \in I \; where, \; I \;=\; \{1, 2, 3, ..., N_s\} \;\}$$

The size of the state space is,

$$|\mathcal{S}| \ = \ 2^{N_B}$$

3.5.4 The MDP Action Space

The MDP action space corresponds to the set of all actions the RL-agent can undertake in a certain state. In the proposed framework, the action space is the combination of SU and frequency band sensing assignments of which the RL-agent picks an optimal action which maximizes the SU-network throughput.

$$\mathcal{A} \ = \ \{ \ \{x_1, x_2, x_3, ..., x_{N_s}\} : \ x_i \in \{0, 1, 2, ..., N_B\} \ \}$$

The size of the action space is,

$$|\mathcal{A}| \ = \ (N_B + 1)^{N_s}$$

3.5.5 Estimation of State Transition Probabilities

In a dynamic radio environment such as a CRN, state transition probabilities will be completely unknown. However, there are methods which allow us to estimate the state transition probabilities. The state transition probabilities may be estimated ONLINE incrementally using the update rule outlined in [4] and the theorem laid down in [5]. Using two consecutive global decisions at the Aggregator node for a specific sub-band, the state transition probabilities denoting the stochastic nature of the dynamics of the radio environment can be estimated.

$$\hat{p}^{n,i}_{jj,k+1} \ = \ \hat{p}^{n,i}_{jj,k} \ + \ \alpha_{p,k}(I_{(d^{n,i}_t = j)} \ - \ \hat{p}^{n,i}_{jj,k}), \ where \ j \in \{0, 1\}$$

Here, k represents the index of the decision pairs in this incremental online algorithm $alpha_{p,k}$ is a step-size parameter in $\{0, 1\}$ $I_{(d^{n,i}_t = j)}$ is an indicator function such that,

$$I_{(d^{n,i}_t)} \ = \ 1, \ if \ (d^{n,i}_t \ = \ j), \ 0 \ otherwise$$

According to the stochastic approximation theorem laid down in [5], [the above update rule converges to the true state transition probabilities if,

$$\sum_{k=0}^{\infty} \alpha_{p,k} \ = \ \infty \ and \ \sum_{k=0}^{\infty} \alpha^2_{p,k} \ < \ \infty, \ i.e. \ for \ instance \ \alpha_{p,k} \ = \ (1/k)$$

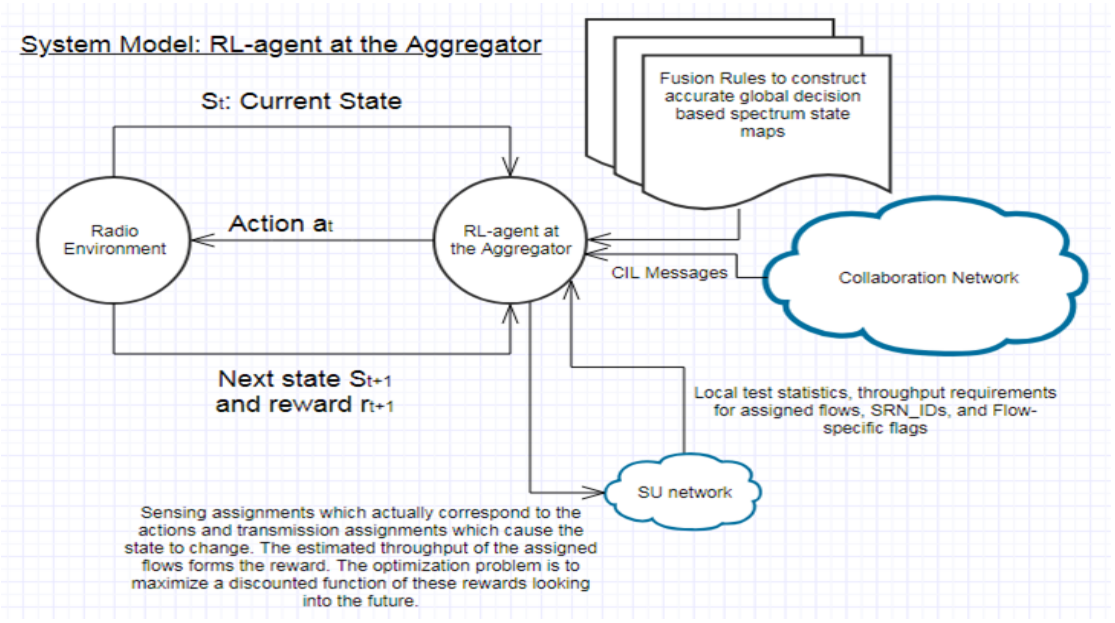

FIGURE 3.4: System Model: Using MDPs at the Aggregator

Figure 3.4 shows the proposed system model employing an MDP-based RL-agent at the Aggregator.

3.5.6 Optimization Problem

The optimization problem involves solving the Optimized Bellman Equation until convergence. The optimization problem can be solved by constructing a Policy Iteration algorithm (Policy Evaluation and Policy Improvement) and solving it using Dynamic Programming (The presence of overlapping sub-problems in the optimization problem, i.e. the recursive relationship between the value function of the current state and the successor state, makes the use of DP optimal). The optimization problem can also be solved by leveraging another construct of DP called Value Iteration which combines policy evaluation and policy improvement into a single update. Asynchronous DP approaches can be used in the above mentioned algorithms to ensure efficient performance in terms of time to convergence, processor memory utilization, and Aggregator energy efficiency. The Finite-Horizon MDP space is given by,

$$\{\mathcal{S}, \mathcal{A}, R, P, \gamma\} \; with \; the \; model \; P(s', r|s, a)$$

Use Asynchronous Dynamic Programming approaches to output the optimal policy $\pi = \pi^*$ such that:

$$\pi(s) = argmax_{a \in \mathcal{A}}\{\sum_{s',r} P(s', r|s, a)[r + \gamma V(s')]\}$$

Policy Iteration (using iterative policy evaluation) for estimating $\pi \approx \pi_*$

1. Initialization
   $V(s) \in \mathbb{R}$ and $\pi(s) \in \mathcal{A}(s)$ arbitrarily for all $s \in \mathcal{S}$

2. Policy Evaluation
   Loop:
   $\Delta \leftarrow 0$
   Loop for each $s \in \mathcal{S}$:
   $v \leftarrow V(s)$
   $V(s) \leftarrow \sum_{s',r} p(s', r|s, \pi(s))[r + \gamma V(s')]$
   $\Delta \leftarrow \max(\Delta, |v - V(s)|)$
   until $\Delta < \theta$ (a small positive number determining the accuracy of estimation)

3. Policy Improvement
   policy-stable $\leftarrow$ *true*
   For each $s \in \mathcal{S}$:
   old-action $\leftarrow \pi(s)$
   $\pi(s) \leftarrow \arg\max_a \sum_{s',r} p(s', r|s, a)[r + \gamma V(s')]$
   If *old-action* $\neq \pi(s)$, then *policy-stable* $\leftarrow$ *false*
   If *policy-stable*, then stop and return $V \approx v_*$ and $\pi \approx \pi_*$; else go to 2

FIGURE 3.5: Solving the Optimization Problem through Policy Iteration

Value Iteration, for estimating $\pi \approx \pi_*$

Algorithm parameter: a small threshold $\theta > 0$ determining accuracy of estimation
Initialize $V(s)$, for all $s \in \mathcal{S}^+$, arbitrarily except that $V(terminal) = 0$

Loop:
  $\Delta \leftarrow 0$
  Loop for each $s \in \mathcal{S}$:
    $v \leftarrow V(s)$
    $V(s) \leftarrow \max_a \sum_{s',r} p(s', r|s, a)[r + \gamma V(s')]$
    $\Delta \leftarrow \max(\Delta, |v - V(s)|)$
until $\Delta < \theta$

Output a deterministic policy, $\pi \approx \pi_*$, such that
  $\pi(s) = \arg\max_a \sum_{s',r} p(s', r|s, a)[r + \gamma V(s')]$

FIGURE 3.6: Solving the Optimization Problem through Value Iteration

Here, $a \in \mathcal{A}$ corresponds to an action in the action space modelled in section 3.5.4, $s \in \mathcal{S}$ corresponds to a state in the state space modelled in section 3.5.3, $P(s', R|s, a)$ refers to the state transition probabilities estimated using stochastic approximation theory, γ is the discount factor such that $\gamma \in [0, 1]$, and $V(s')$ refers to the value function estimate of the successor state in the previous evaluation cycle. Figure 3.5 shows the algorithm to solve the Optimization Problem termed Policy Iteration (Reference 6). Figure 3.6 shows the algorithm to solve the Optimization Problem termed Value Iteration (Reference 6).

3.6 Model-Free learning

3.6.1 Key Points

- Using the same State Space and Action Space modelling outlined in the previous approach, learn the model of the radio environment using deep searches / sampled searches / n-step deep weighted searches over the back-off diagram.
- Monte-Carlo Learning: Deep and Sampled backup (Learn directly from episodes of experience)
- Temporal Difference Learning (TD()): Online update of value functions even after incomplete sequences, introduction of eligibility traces, combining n-step returns using weights, incremental approach more efficient than MC learning.

3.6.2 Incremental Monte Carlo Learning

Monte-Carlo Learning is a model-free learning technique that does not require any knowledge about the state transition probabilities of the dynamic system under observation. The goal here is to learn the value function v_π from episodes of experience under π.

Update $v_\pi(s)$ incrementally after each episode under π: $S_i,\ A_i,\ R_{i+1},\ S_{i+1}$. For each state $s \in \mathcal{S}$ with return G_t,

$$N(s_t) \ \longleftarrow \ N(s_t) + 1$$

$$v_\pi(s_t) \ \longleftarrow \ v_\pi(s_t) \ + \ ((G_t - v_\pi(s_t))/N(s_t))$$

3.6.3 Incremental Online TD(λ) Learning

Update the value function under π for each state S$in\mathcal{S}$ as follows,

$$v_\pi(s) \ = \ v_\pi(s) \ + \ \alpha\delta_t E_t(s)$$

where, $E_t(s)$ refers to the Eligibility trace of state s, α is the step-size parameter, and δ_t is the TD error such that,

$$\delta_t \ = \ r_{t+1} \ + \ \gamma v(s_{t+1}) \ - \ v(s_t)$$

Collaboration Container Interface	col0
Collaboration Network	172.30.<CollabNet>.<HostID> / 24
CollabNet	101-228
Host IDs (for SRNs 1 128)	101-228
Collaboration Server IPs	172.30.<CollabNet>.2 / 24

TABLE 3.1: Collaboration Network Design Parameters

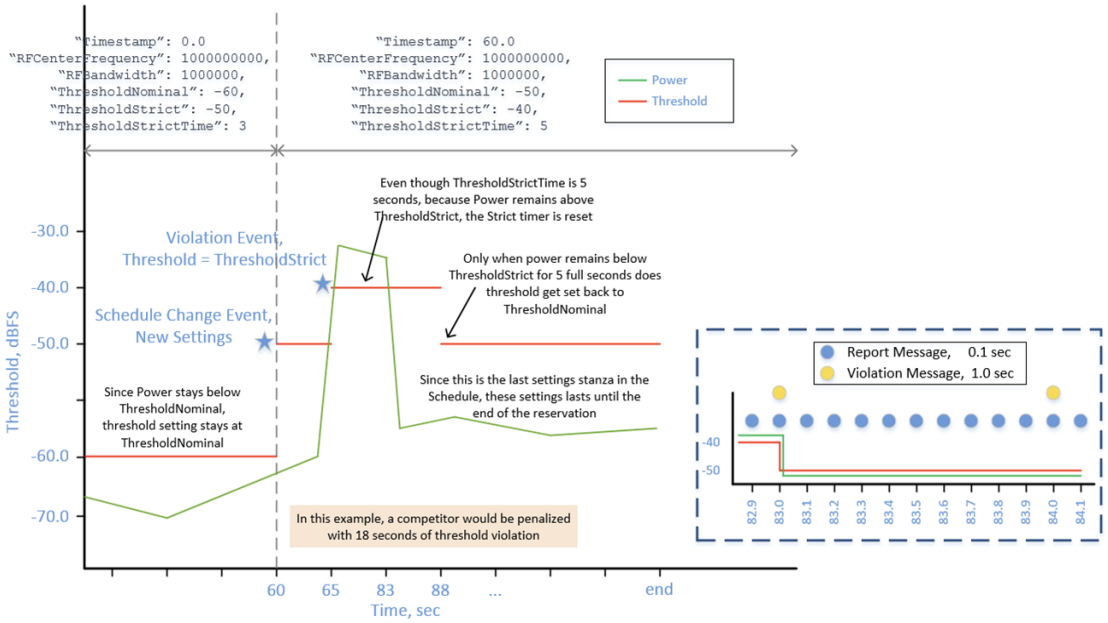

FIGURE 3.7: DARPA Passive Incumbent Message Specification

3.7 DARPA SC2 Collaboration Network Design

A collaboration network and server will be allocated to each SC2 reservation. The network and server will only be accessible by nodes within the reservation. The collaboration network will consist of a single /24 broadcast domain and will only require link local host routes. The IP address 3rd octet will be the same for all nodes and server in a reservation and will be defined at allocation time. Each collaboration gateway interface will be named col0. The servers IP address 4th octet will always be 2. Table 3.1 summarizes the network design for the SC2 Collaboration System. A configuration file will be pushed into each competitor container that provides the IP address of the collaboration server. For information on this file, see the colosseum_config.inisection of the Radio Command and Control (C2) API specification on FreshDesk.

Figure 3.7 shows illustrates a scenario highlighting the IncumbentNotify CIL message specifications corresponding to the Passive Incumbent in the Collaboration Network (Source: DARPA SC2 Colosseum FreshDesk)

3.7.1 Modelling the CIL messages into the framework (common for all approaches)

- Incorporate the Report and Violation Messages from the Collaboration Server into the reward distribution heuristics. If a Violation is reported, penalize the sub-band and assignment pair state value function or action value function by the number of seconds of violation reported in the Violation message.
- Use the parameters in the Report Message and scheduled parameter change messages to modify the detection threshold / constraint vectors in the Aggregators fusion rules.
- Use other CIL messages such as the SpectrumUsageMessage to get a more well-informed global view of spectrum utilization by using simple OR or K-out-of-N fusion heuristics at the Aggregator.

Chapter 4

Distributed Collaborative Cognitive Radio Environments with Neighbour Discovery

As of this version (v1.0.0) of the thesis proposal, only Phase 1 has been investigated in detail and will be taken up during the first semester of the Master of Science degree with the phase ending in its successful incorporation in the BAM! Wireless code-base and one or more IEEE research publications. This phase is presented as a potential candidate which can be tackled in semesters II and beyond.

4.1 Overview

One proposal is to employ Clustering algorithms to produce highly correlated clusters of collaborating neighbours. These neighbours may then proceed to nominate a cluster-head which would operate as another standard radio node in the next tier with other cluster-heads and serve as a centralized service node for its cluster. The cluster-heads would have Reinforcement Learning agents running on them which try to optimize the formulated problem and come up with an optimal or a nearly-optimal action policy that maximizes the SU network throughput while imposing constraints on missed detection probability, control channel overhead, energy efficiency of nodes, and decision latencies.

Another approach here would be to employ multi-agent, multi-band, distributed Reinforcement Learning agents on individual SUs. The proposal as of this version of the document is to replicate the results obtained in 4.

A third approach would be replicate the results detailed in 8 which employs Markov Decision Processes at SUs to learn their optimal set of cooperating neighbours in CRAHNs. This proposal is described to minimize control channel overhead and improve detection performance by using binary decision dissemination and hard-combining strategy.

Chapter 5

Opportunistic spectrum access in competing radio resource utilization environments

As of this version (v1.0.0) of the thesis proposal, only Phase 1 has been investigated in detail and will be taken up during the first semester of the Master of Science degree with the phase ending in its successful incorporation in the BAM! Wireless code-base and one or more IEEE research publications. This phase is presented as a potential candidate which can be tackled in semesters II and beyond.

5.1 Overview

The proposal aims to use Bandit frameworks and Reinforcement Learning algorithms to devise optimum action policies for sensing & transmission assignments and sensing & transmission schedules. As outlined in the introductory sections of the document, non-collaborative spectrum access environments require novel, intelligent back-off strategies along with mandated acknowledgements (with some class of ARQ) in order to facilitate reliable completion of assigned network flows in a competing radio environment where there is no knowledge transfer among radio nodes. Also, in these environments, the problems of fading and shadowing that typically creep into PU activity measurements at the individual, non-collaborating SU, have to be solved.

Chapter 6

Detailed Performance evaluations of the proposed heuristics in specific topology/traffic scenarios

As of this version (v1.0.0) of the thesis proposal, only Phase 1 has been investigated in detail and will be taken up during the first semester of the Master of Science degree with the phase ending in its successful incorporation in the BAM! Wireless code-base and one or more IEEE research publications. This phase is presented as a potential candidate which can be tackled in semesters II and beyond.

6.1 Overview

The following are potential categories of research which can be undertaken in Phase 4.

- The incorporation of autonomous participation strategies at each SU in the network leveraging the cost-reward trade-off analyses in Bayesian games. Intuitively, the SU decides to either participate in the collaboration process or to go at it alone by doing a trade-off analyses with respect to the obtained reward and the paid cost. In the simplest terms, the cost for an SU to participate in a collaborative sensing episode with its neighbours would be the time lost sensing frequency bands which could have been used for transmission/sensing other successful frequency bands and the reward would be the throughput of the successfully completed flow over the channel found to be idle as a result of this collaborative spectrum sensing episode.

- Sensing and Transmission scheduling heuristics as described in 12
- Determining PU activity using supervised learning algorithms, for instance, determining PU channel access schemes using CNNs/SVMs. Another approach to determine PU channel access schemes would be to use cumulant based classifiers as detailed in 10
- Using stage-based, iterative addition/merger/deletion/updation of PU components (these so-called components model the PU footprint over the region of interest) by using multi-stage aggregation. One approach would be to use the soft and hard reports heuristics described in 11.
- Detailed Performance evaluation of these heuristics against the algorithms developed in Phase 1, 2, and 3 by varying numerous design parameters such as the diversity order of sensing, the heterogeneity of nodes, the number of bands sensed per node, different reward distributions, with or without prior knowledge of the network, with or without CSI, and many other control knobs which may turn up as we proceed with the design stages of phases 1, 2, and 3.

Chapter 7

Coalescence of various approaches into a best-policy framework and possible applications in the 5G and IoT landscape

As of this version (v1.0.0) of the thesis proposal, only Phase 1 has been investigated in detail and will be taken up during the first semester of the Master of Science degree with the phase ending in its successful incorporation in the BAM! Wireless code-base and one or more IEEE research publications. This phase is presented as a potential candidate which can be tackled in semesters III and beyond.

7.1 Overview

- The consolidation of all the knowledge obtained in phases 1, 2, 3, and 4 into a "best-policy" framework in which the SU network adapts to changing circumstances by learning the topology of the network it's operating in and choosing the best possible operational action heuristic to maximize the SU network throughput in collaborative topologies or to successfully complete its assigned traffic flows in non-collaborative or opportunistic topologies.
- The SU node and SU network adaptation can be approached coarsely leveraging the programmability brought forth by Software Defined Networking. Simple protocols like REST and CLI can be employed to modify SU behaviour with respect

to changing network topologies and flow priorities by automating parameter variations triggered at the application layer which then trickle down to the data plane through the control plane.

- A fine-grained approach would be to embed this intelligence into the SU network and modify SU behaviour with variations in network topology, flow priorities, and other secondary design considerations.

- This "best-policy" framework would be prototyped on SDR and GPU test-beds to complete the design flow. The vector processing and multi-threading capabilities of GPUs can be leveraged by allocating data-intensive tasks such as model training to the GPUs. The system's software architectural choices such as type of PUB-SUB framework used, design patterns, asynchronous calls, design patterns, dedicated I/O services, loosely-coupled design strategies, and allocation of operations across FPGAs, GPUs, and CPUs play a vital role in the overall performance of the proposed framework. Hence, E2E prototyping is a required phase.

- Furthermore, potential applications of the framework and the algorithms under its hood, in the 5G and the IoT landscape can be explored.

Chapter 8

Conclusion

The graduate research proposal detailed in this document aims to devise a novel, adaptive, hierarchical, intelligent framework for Secondary Users in Cognitive Radio Networks leveraging algorithms and heuristics from Reinforcement Learning, Bandits, Supervised/Unsupervised Learning, and Game Theory. The proposed research endeavour is broken down into five phases with each phase ending in one or more IEEE research publications and potential inclusion in the BAM! Wireless code-base for emulation on the DARPA SC2 Colosseum. The end goal of this research is to produce the over-arching framework that's been extensively discussed in this document and explore its potential applications in fifth-generation wireless networks and the IoT ecosystem where Massive MTC demands intelligent dynamic spectrum access solutions.

Bibliography

[1] J. Oksanen, V. Koivunen, J. Lundn and A. Huttunen, "Diversity-based spectrum sensing policy for detecting primary signals over multiple frequency bands," 2010 IEEE International Conference on Acoustics, Speech and Signal Processing, Dallas, TX, 2010, pp. 3130-3133.

[2] J. Oksanen, V. Koivunen, J. Lundn, A. Huttunen, Diversitybased Spectrum Sensing Policy for Detecting Primary Signal Over Multiple Frequency Bands, in: Proc. of the ICASSP Conference, Dallas Texas, 31303133, 2010.

[3] S. Chaudhari, V. Koivunen, H. V. Poor, Autocorrelation-Based Decentralized Sequential Detection of OFDM Signals in Cognitive Radios, IEEE Trans. Signal Process. 57 (7) (2009) 2690 2700.

[4] J. Lundn, S. R. Kulkarni, V. Koivunen and H. V. Poor, "Multiagent Reinforcement Learning Based Spectrum Sensing Policies for Cognitive Radio Networks," in IEEE Journal of Selected Topics in Signal Processing, vol. 7, no. 5, pp. 858-868, Oct. 2013, doi: 10.1109/JSTSP.2013.2259797.

[5] T. Jaakkola, M. I. Jordan, and S. P. Singh, On the convergence of stochastic iterative dynamic programming algorithms, Neural Comput., vol. 6, pp. 11851201, 1994.

[6] R. S. Sutton, A. G. Barto, Reinforcement Learning: An Introduction, Cambridge, MA: MIT Press, 2018

[7] J. Lundn, V. Koivunen, S. R. Kulkarni and H. V. Poor, "Exploiting spatial diversity in multiagent reinforcement learning based spectrum sensing," 2011 4th IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP), San Juan, 2011, pp. 325-328, doi: 10.1109/CAMSAP.2011.6136016.

[8] B. F. Lo and I. F. Akyildiz, "Reinforcement learning-based cooperative sensing in cognitive radio ad hoc networks," 21st Annual IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, Instanbul, 2010, pp. 2244-2249, doi: 10.1109/PIMRC.2010.5671686.

[9] A. Anandkumar, N. Michael and A. Tang, "Opportunistic Spectrum Access with Multiple Users: Learning under Competition," 2010 Proceedings IEEE INFOCOM, San Diego, CA, 2010, pp. 1-9, doi: 10.1109/INFCOM.2010.5462144.

[10] M. Laghate, P. Urriza and D. Cabric, "Channel Access Method Classification for Cognitive Radio Applications," in IEEE Wireless Communications Letters, vol. 7, no. 1, pp. 70-73, Feb. 2018, doi: 10.1109/LWC.2017.2754367.

[11] M. Laghate and D. Cabric, "Cooperatively Learning Footprints of Multiple Incumbent Transmitters by Using Cognitive Radio Networks," inIEEE Transactions on Cognitive Communications and Networking, vol. 3, no. 3, pp. 282-297, Sept. 2017. doi: 10.1109/TCCN.2017.2710309.

[12] A. Azarfar, C. Liu, J. Frigon, B. Sans and D. Cabric, "Joint transmission and cooperative spectrum sensing scheduling optimization in multi-channel dynamic spectrum access networks,"2017 IEEE International Symposium on Dynamic Spectrum Access Networks (DySPAN), Piscataway, NJ, 2017, pp.1-10, doi: 10.1109/DySPAN.2017.7920789.

[13] M. Laghate and D. Cabric, "Using multiple power spectrum measurements to sense signals with partial spectral overlap," 2017 IEEE International Symposium on Dynamic Spectrum Access Networks (DySPAN), Piscataway, NJ, 2017, pp. 1-8, doi: 10.1109/DySPAN.2017.7920751.

[14] A. Das, S. C. Ghosh, N. Das and A. D. Barman, "Q-Learning Based Co-Operative Spectrum Mobility in Cognitive Radio Networks," 2017 IEEE 42nd Conference on Local Computer Networks (LCN), Singapore, 2017, pp. 502-505, doi: 10.1109/LCN.2017.80.

[15] M. H. Hassan, M. J. Hossain and V. K. Bhargava, "Distributed Beamforming and Autonomous Participation Decision Making in Cooperative CR Systems in Presence of Asynchronous Interference," in IEEE Transactions on Wireless Communications, vol. 15, no. 7, pp. 5016-5029, July 2016, doi: 10.1109/TWC.2016.2551219.

[16] X. Jiang and H. Xi, "A POMDP approach to opportunistic spectrum access with feedback," Proceedings of the 11th IEEE International Conference on Networking, Sensing and Control, Miami, FL, 2014, pp. 690-694, doi: 10.1109/ICNSC.2014.6819709.

[17] S. Hu, Y.-D. Yao, and Z. Yang, MAC protocol identification using support vector machines for cognitive radio networks, IEEE Wirel. Commun., vol. 21, no. 1, pp. 5260, Feb. 2014.

[18] S. A. Rajab, W. Balid, M. O. A. Kalaa, and H. H. Refai, Energy detection and machine learning for the identification of wireless MAC technologies, in Proc. IWCMC, Aug. 2015, pp. 14401446.