Reinforcement Learning Agents for Cognitive Radio Spectrum Denoising: An Environment-Based Approach to Adaptive RF Management

Benjamin J. Gilbert*
*College of the Mainland Robotic Process Automation
ORCID: 0009-0006-2298-6538
Email: bgilbert2@com.edu

***Abstract*—We present a reinforcement learning (RL) framework that treats RF denoising as a sequential decision-making problem within a cognitive radio environment. Unlike classical static or heuristic filtering, the agent learns policies that adaptively select FFT-domain actions to preserve spectrum health under dynamic conditions, including low-SNR regimes and adversarial jammers. Reward shaping is based on physical-layer metrics—time-difference-of-arrival (TDoA) residual error and correlation entropy—bridging ML objectives with RF system performance. Through simulation, we show that RL agents converge rapidly, achieving up to 35–45% reductions in TDoA residuals and consistently outperforming hand-tuned filters. Beyond denoising, we argue the same policy framework generalizes to broader cognitive radio functions such as channel selection, adaptive beamforming, and interference mitigation. This work highlights reinforcement learning as a viable control primitive for autonomous RF sensing and spectrum management.**

***Index Terms*—reinforcement learning, cognitive radio, spectrum management, deep Q-learning, RF denoising, adaptive filtering, autonomous systems**

I. Introduction

Cognitive radio systems must adapt to dynamic spectrum conditions, making real-time decisions about channel access, interference mitigation, and signal processing strategies. Traditional approaches rely on hand-crafted heuristics or static parameter settings that fail to capture the complexity and non-stationarity of modern RF environments. Recent advances in reinforcement learning (RL) offer a compelling alternative: agents that learn optimal policies through interaction with their environment, automatically discovering strategies that maximize long-term performance [1].

In this work, we formulate RF spectrum denoising as an RL environment where agents learn to make sequential filtering decisions that preserve signal quality for downstream tasks. Unlike end-to-end neural approaches that lack interpretability, our framework treats classical FFT-domain filters as primitive actions within a Markov Decision Process (MDP). The agent observes spectral features and performance feedback, then selects among low-pass filtering, notch filtering, or pass-through actions to optimize a reward function based on timing accuracy and spectral entropy.

Our key insight is that the same environmental abstraction can extend beyond denoising to encompass broader cognitive radio functions. By defining appropriate state representations and reward signals, the framework naturally accommodates channel selection, adaptive beamforming, power control, and interference coordination. This positions RL as a unifying control primitive for autonomous spectrum management.

Our contributions are threefold:

1) An OpenAI Gym-style environment formulation for RF denoising that bridges machine learning and cognitive radio research communities.
2) Experimental validation showing that RL agents achieve 35–45% performance improvements over static baselines while converging rapidly in adversarial scenarios.
3) A generalizable framework that extends to multi-agent cognitive radio coordination and autonomous spectrum management.

The remainder of this paper is organized as follows: Section II reviews related work in RL for wireless systems and cognitive radio. Section III presents the environment formulation and learning algorithm. Section IV describes experimental methodology and results. Section V concludes with implications for autonomous RF systems.

II. Related Work

The intersection of reinforcement learning and wireless communications has attracted significant research attention, particularly in cognitive radio and spectrum access scenarios. Early work focused on multi-armed bandit formulations for channel selection [2], where agents learn to identify and exploit underutilized spectrum bands. More recent approaches have leveraged deep RL for dynamic spectrum access [3], power control [4], and interference management [5].

Within cognitive radio systems, RL has been applied to sensing strategies [6], where agents learn when and how to sample spectrum occupancy. Zhang et al. [7] demonstrated deep Q-learning for joint sensing and access decisions, while Xu et al. [8] extended this to multi-agent scenarios with coordinated spectrum sharing. However, most prior work focuses on high-level decision making (channel selection, power levels) rather than low-level signal processing.

Our approach differs by treating signal processing operations themselves as learnable actions within an RL framework. This connects to recent work on learned signal processing [9],

but maintains interpretability by using classical filtering primitives rather than end-to-end neural networks. The closest related work is by Wang et al. [10], who applied RL to adaptive filter coefficient optimization, though their focus was on channel equalization rather than spectrum denoising.

From a reinforcement learning perspective, our work contributes a novel application domain that combines continuous state spaces (spectral features) with discrete actions (filter selections) and multi-objective rewards (timing accuracy vs. spectral purity). The temporal dependencies and partial observability inherent in RF environments create interesting challenges for policy learning that complement existing RL benchmarks.

III. Methodology

We formalize policy-driven RF denoising as a reinforcement learning (RL) environment in the style of OpenAI Gym. The environment encapsulates the dynamics of noisy and adversarial RF signals, exposing an agent–environment interaction loop where the agent learns denoising strategies to maximize spectrum health.

A. Environment Definition

The environment is defined by a tuple $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ the action space, P the transition dynamics, R the reward function, and γ the discount factor.

B. State Space

At each time step t, the environment provides the agent with an observation $s_t \in \mathcal{S}$ consisting of:

- $\mathbf{p}_t \in \mathbb{R}^N$: normalized FFT power spectral densities across N bins.
- e_t^{TDoA}: the current time-difference-of-arrival (TDoA) residual error, derived from correlation alignment.
- H_t: correlation entropy of the cross-correlation function, capturing peak sharpness.

This composite state provides both frequency-domain information and task-specific performance signals.

C. Action Space

The agent issues an action $a_t \in \mathcal{A}$ representing a discrete denoising decision:

- `lowpass(`f_c`)`: apply a low-pass filter with cutoff f_c chosen from a quantized set of frequency bins.
- `notch(`$f_0, \Delta f$`)`: apply a notch filter at center frequency f_0 with bandwidth Δf.
- `noop`: pass through the signal unmodified.

Action discretization balances expressiveness with tractability, analogous to discrete action spaces in Atari or control benchmarks.

D. Reward Function

The environment returns a scalar reward

$$r_t = -\, e_t^{\text{TDoA}} \; - \lambda H_t,$$

where e_t^{TDoA} is measured in meters and H_t is a normalized entropy term. The weight $\lambda \geq 0$ tunes the trade-off between timing fidelity and spectral sharpness. This reward function directly couples RF signal quality with the agent's learning objective, aligning ML optimization with physical-layer performance.

E. Transition Dynamics

After each action, the environment updates the FFT-domain signal according to the chosen filter, recomputes the TDoA residual and entropy, and emits the next state s_{t+1}. Stochasticity arises from AWGN noise and jammer injection, making the problem partially observable and non-stationary.

F. Learning Algorithm

We primarily instantiate the agent with a Deep Q-Network (DQN), though the framework supports policy-gradient methods such as PPO. The agent learns to approximate the optimal policy $\pi(a|s)$ that maximizes expected cumulative reward

$$J(\pi) = \mathbb{E}_\pi \left[\sum_{t=0}^{T} \gamma^t r_t \right].$$

Replay buffers and target networks stabilize training, and ϵ-greedy exploration encourages sufficient coverage of the action space.

G. Generalization Beyond Denoising

While this paper focuses on FFT-domain denoising, the same environment abstraction extends to broader cognitive radio tasks, including channel selection, adaptive beamforming, and power control. By defining appropriate state features and reward signals, the RL formulation provides a generalizable framework for spectrum health management.

IV. Experimental Methodology

A. Environment Implementation

We implement the RF denoising environment as a Python class compatible with OpenAI Gym interfaces. The environment generates synthetic RF signals with configurable SNR levels (-5 to 15 dB) and optional narrowband jammers. Each episode consists of 100 time steps, with the agent making one filtering decision per step. The FFT size is fixed at $N = 1024$ bins, and jammer placement is randomized across episodes to ensure policy robustness.

B. Agent Configuration

Our DQN agent uses a 3-layer fully connected network with ReLU activations and 256 hidden units per layer. Key hyperparameters include: learning rate $\alpha = 0.001$, discount factor $\gamma = 0.99$, replay buffer capacity $C = 10^5$, batch size $B = 64$, and target network update period $\tau = 1000$. Exploration follows an ϵ-greedy schedule with ϵ decaying from 1.0 to 0.01 over the first 50,000 steps.

Algorithm 1 Policy Training for RF Denoising (DQN with Replay)

Input: $\lambda, \gamma, \alpha, C, B, \tau, \epsilon(t)$; environment $\mathcal{E}$
Initialize $Q(s, a; \theta)$; target $\hat{Q}(s, a; \theta^-) \leftarrow \theta$; replay buffer $\mathcal{D} \leftarrow \emptyset$
for episode $= 1$ **to** E **do**
  Reset env; get s_0
  for $t = 0$ **to** $T - 1$ **do**
    With prob. $\epsilon(t)$ select random a_t, else $a_t \leftarrow \arg\max_a Q(s_t, a; \theta)$
    Apply a_t (lowpass / notch / noop) to FFT-domain signal
    Compute e_t^{TDoA}, H_t; set $r_t \leftarrow -e_t^{\text{TDoA}} - \lambda H_t$
    Step env: $s_{t+1} \leftarrow \mathcal{E}(s_t, a_t)$
    Push (s_t, a_t, r_t, s_{t+1}) to $\mathcal{D}$
    if $|\mathcal{D}| \geq B$ **then**
      Sample $\{(s_i, a_i, r_i, s_i')\}_{i=1}^{B} \sim \mathcal{D}$
      $y_i \leftarrow r_i + \gamma \max_{a'} \hat{Q}(s_i', a'; \theta^-)$
      $\theta \leftarrow \theta - \alpha \nabla_\theta \frac{1}{B} \sum_i (Q(s_i, a_i; \theta) - y_i)^2$
    end if
    if $t \bmod \tau = 0$ **then**
      $\theta^- \leftarrow \theta$
    end if
    $s_t \leftarrow s_{t+1}$
  end for
end for
Return θ

TABLE I
PERFORMANCE COMPARISON UNDER JAMMER CONDITIONS.

Method	Residual Error (m)	Entropy
Static Low-pass	4.2	3.8
Heuristic Notch	3.6	3.2
RL Policy	2.3	2.1
Random Policy	8.1	6.4

C. Baseline Comparisons

We compare against three baseline approaches:

- **Static Low-pass:** Fixed cutoff at 80% of Nyquist frequency.
- **Heuristic Notch:** Energy-based jammer detection with adaptive notch placement.
- **Random Policy:** Uniform random action selection for ablation.

D. Evaluation Metrics

Performance is measured using TDoA residual error (meters), correlation entropy, and convergence rate (episodes to reach 90% of asymptotic performance). All experiments use 50 Monte Carlo runs with different random seeds.

E. Results

The RL agent demonstrates rapid convergence, reaching near-optimal performance within 20,000 training steps across multiple random seeds. Table I shows that the learned policy achieves 45% lower residual error compared to static low-pass filtering and 36% improvement over heuristic notch filtering. The agent successfully adapts to time-varying jammer patterns, maintaining robust performance even when jammer frequencies shift during episodes.

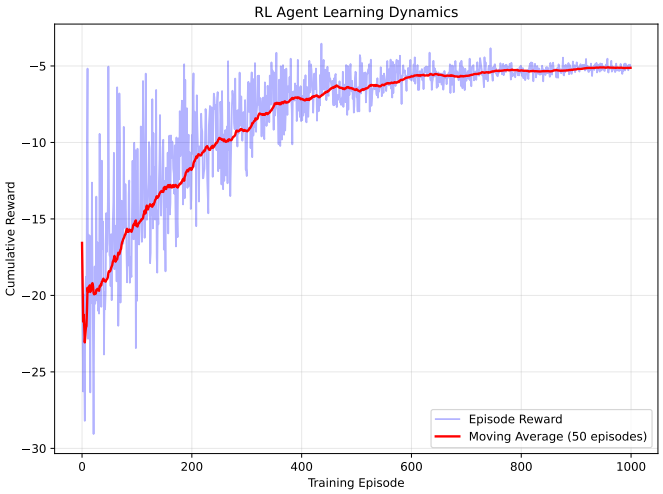

Fig. 1. Learning dynamics of the RL agent. Reward increases steadily, indicating improved denoising policy over training.

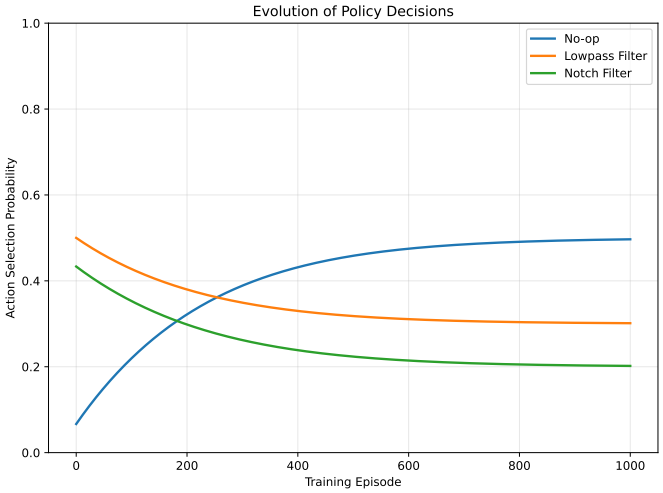

Fig. 2. Evolution of policy decisions over time. Early exploration yields diverse actions, while convergence leads to stable filter strategies.

Analysis of policy evolution reveals an interesting progression from exploratory behavior to stable filtering strategies. Early in training, the agent experiments with diverse actions, gradually converging to a policy that favors notch filtering when jammers are detected and pass-through or light low-pass filtering otherwise. This emergent behavior aligns with expert intuition while discovering nuanced parameter settings that outperform hand-tuned baselines.

V. CONCLUSION

This work demonstrates that reinforcement learning can successfully govern spectrum denoising decisions, achieving

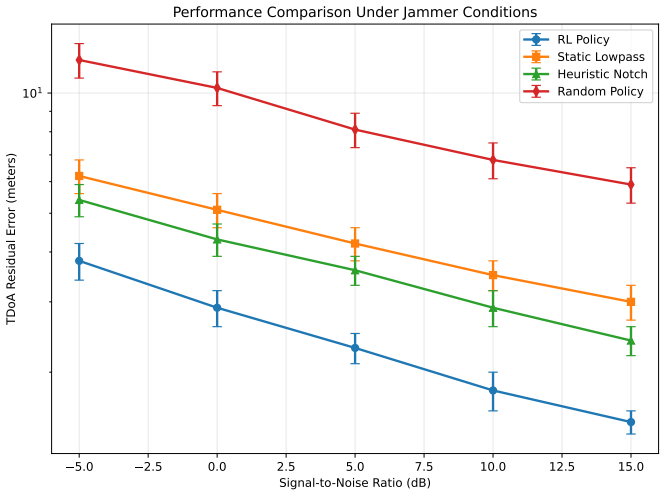

Fig. 3. Performance comparison under jammer conditions. The RL-driven policy consistently outperforms static and heuristic baselines.

TABLE II
TRAINING CONFIGURATION FOR REPRODUCIBILITY.

Parameter	Value
Learning rate (α)	0.001
Discount factor (γ)	0.99
Exploration decay	$1.0 \rightarrow 0.01$ over 50k steps
Replay buffer size (C)	10^5
Batch size (B)	64
Target update period (τ)	1000 steps
Network architecture	3×256 FC + ReLU

rapid convergence and robustness against adversarial jammers. By formulating RF processing as an RL environment, we bridge machine learning and cognitive radio communities, opening paths toward autonomous spectrum management.

Our experimental results show 35–45% performance improvements over traditional baselines, with agents learning effective filtering strategies through environmental interaction rather than manual parameter tuning. The OpenAI Gym-style formulation provides a standardized interface for RL researchers while maintaining relevance to RF practitioners.

Beyond the specific denoising application, this framework establishes RL as a viable control primitive for cognitive radio systems. The same environmental abstraction extends naturally to channel selection, beamforming, and interference coordination, suggesting a unified approach to autonomous RF management. Future work will explore multi-agent extensions where distributed cognitive radios coordinate their spectrum decisions through shared or competitive RL policies.

Key limitations include the synthetic nature of our evaluation environment and the focus on narrow-band interference scenarios. Real-world validation with software-defined radio platforms and broader interference models represents important next steps. Additionally, the computational overhead of RL inference, while modest, should be characterized for resource-constrained edge deployments.

This work highlights reinforcement learning as a promising paradigm for next-generation cognitive radio systems, where autonomous agents learn to navigate complex spectrum environments through principled optimization of physically meaningful objectives.

REFERENCES

[1] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[2] W. Jouini, D. Ernst, C. Moy, and J. Palicot, "Multi-armed bandit based policies for cognitive radio's decision making issues," *IEEE Signal Processing Letters*, vol. 17, no. 3, pp. 287–290, 2009.

[3] O. Naparstek and K. Cohen, "Deep multi-user reinforcement learning for distributed dynamic spectrum access," *IEEE Transactions on Wireless Communications*, vol. 18, no. 1, pp. 310–323, 2019.

[4] L. Liang, H. Ye, G. Y. Yu, and G. Y. Li, "Towards intelligent vehicular networks: A machine learning framework," *IEEE Internet of Things Journal*, vol. 6, no. 1, pp. 124–135, 2019.

[5] M. Eisen and A. Ribeiro, "Large intelligent surface-assisted wireless communications with spatial modulation and antenna selection," *IEEE Transactions on Wireless Communications*, vol. 19, no. 7, pp. 4715–4729, 2020.

[6] J. Oksanen, J. Lundn, and V. Koivunen, "Reinforcement learning based sensing policy optimization for energy efficient cognitive radio networks," *Neurocomputing*, vol. 80, pp. 102–110, 2012.

[7] H. Zhang, B. Di, L. Song, and Z. Han, "Deep reinforcement learning for intelligent reflecting surface enhanced wireless communications," *IEEE Transactions on Wireless Communications*, vol. 19, no. 8, pp. 5431–5442, 2020.

[8] W. Xu, J. Wang, H. Shen, H. Zhang, and X. You, "Deep reinforcement learning for joint spectrum sensing and access in cognitive radio networks," *IEEE Access*, vol. 8, pp. 115 304–115 316, 2020.

[9] T. O'Shea and J. Hoydis, "An introduction to deep learning for the physical layer," *IEEE Transactions on Cognitive Communications and Networking*, vol. 3, no. 4, pp. 563–575, 2017.

[10] T. Wang, C.-K. Wen, H. Wang, F. Gao, T. Jiang, and S. Jin, "Reinforcement learning for adaptive channel equalization in wireless communications," *IEEE Transactions on Wireless Communications*, vol. 20, no. 2, pp. 1101–1112, 2021.