**Interkoneksi**
**Journal of Computer Science and Digital Business**
ISSN: 3031-3910 Vol. 4 No. 1 (2026)
DOI: https://doi.org/10.61166/interkoneksi.v4i1.95 pp. 208-222

Research Article

# A Stackelberg Game-Theoretic Framework with Q-Learning-Based Adaptive Thresholding for Mitigating Primary User Emulation Attacks in OFDM-Based Cognitive Radio Networks

**Mohsin Mahmood[1], AbdulBasir Momand[2], Abdul Satar Popalzai[3], Sohaib Ahmad Khalil[4], Tauseef Alam Qureshi[1]**

1. Faculty of Computing and Numerical Sciences, City University of Science and Information Technology, Peshawar, Pakistan.
2. Faculty of Computer Science, Rana University, Kabul, Afghanistan
3. Information Technology Department, Faculty of Computer Science, Hewad University, kabul, Afghanistan.
4. Department of Computer Science, Abasyn University, Peshawar, Pakistan

Corresponding Author, E-mail; enmohsing@gmail.com (Mohsin Mahmood)



Received : June 19, 2026 Revised : July 16, 2026
Accepted : August 13, 2026 Available online : September 05, 2026

**How to Cite:** Mohsin Mahmood, Abdul Basir Momand, Abdul Satar Popalzai, Sohaib Ahmad Khalil, & Tauseef Alam Qureshi. (2026). A Stackelberg Game-Theoretic Framework with Q-Learning-Based Adaptive Thresholding for Mitigating Primary User Emulation Attacks in OFDM-Based Cognitive Radio Networks. *Interkoneksi: Journal of Computer Science and Digital Business*, *4*(1), 208–222. https://doi.org/10.61166/interkoneksi.v4i1.95

**Abstract.** Primary User Emulation Attack (PUEA) is a critical denial-of-service threat in OFDM-based Cognitive Radio Networks (CRNs), in which an adversary mimics the signal characteristics of a licensed

primary user (PU) to deny legitimate secondary users (SUs) access to idle spectrum. Conventional energy-detection sensing relies on a *fixed* decision threshold and is therefore structurally unable to adapt to a strategic, power-adjusting attacker. This paper develops a game-theoretic defense framework that models the interaction between the cognitive radio network (CRN) — acting as a Stackelberg *leader* that sets the spectrum-sensing threshold — and the PUEA attacker — acting as a *follower* that chooses its emulation transmit power. We show that the attacker's optimization is ill-posed under a detection-probability-only cost term, since attack success can be driven arbitrarily close to unity by unbounded transmit power, and we resolve this by introducing an exposure/power cost into the attacker's utility, yielding a well-defined best response. We derive, in closed form, the attacker's optimal emulation power as a function of the sensing threshold, and show that at the resulting equilibrium the attack success probability becomes *invariant* to the CRN's threshold choice for a fixed exposure cost — a structural property of the PUEA game that, to our knowledge, has not been reported in the existing literature. Building on this result and an augmented CRN utility that explicitly penalizes missed detection of the true PU, we derive a closed-form, Neyman–Pearson-type expression for the equilibrium sensing threshold. To relax the requirement that the CRN know the attacker's cost and channel parameters exactly, we embed the closed-form equilibrium as a warm start for a Q-learning agent that adaptively refines the threshold from observed sensing outcomes. We present the full mathematical derivation, an algorithmic realization of the combined Stackelberg–Q-learning scheme, a complexity and convergence discussion, and a numerical evaluation of the closed-form expressions across representative parameter ranges that confirms the predicted monotonic trends and reproduces classical asymptotic detection-theoretic behavior as a consistency check. Full OFDM physical-layer Monte Carlo and hardware testbed validation of the learning component are identified as the immediate next stage of this work. The proposed framework gives CRN operators a principled, adaptive alternative to fixed-threshold spectrum sensing under strategic PUEA.

**Keywords:** Cognitive radio networks; primary user emulation attack; spectrum sensing; Stackelberg game; Nash equilibrium; energy detection; Q-learning; reinforcement learning; OFDM; cybersecurity.

## INTRODUCTION

Cognitive radio (CR) was proposed as a means of alleviating spectrum scarcity by allowing unlicensed secondary users (SUs) to opportunistically access frequency bands left idle by licensed primary users (PUs) [1]. The CR paradigm rests on a simple contract: an SU may transmit only when it has reliably established, through spectrum sensing, that the band is unoccupied by the PU. This dependence on sensing is also the paradigm's principal security weakness. In a Primary User Emulation Attack (PUEA), a malicious node transmits a signal that reproduces the statistical or waveform characteristics of the legitimate PU, causing SUs to (incorrectly) vacate or avoid the channel [2],[3]. Because PUEA requires no compromise of cryptographic keys, protocol state, or higher-layer identifiers, it is difficult to counter with conventional authentication mechanisms, and it can be mounted with commodity software-defined radio hardware.

The dominant family of countermeasures in the literature detects PUEA through localization or transmitter fingerprinting [10], through cooperative and Bayesian belief-updating sensing [4],[5], or by treating the interaction between the SU network and the attacker as a game and solving for an equilibrium sensing/attack policy [5]–[9]. The game-theoretic line of work is attractive because PUEA is, at its core, an adversarial and adaptive process: a static, fixed-threshold detector can always

be probed and exploited by an attacker who adjusts its transmit power to sit just below the detector's boundary. A game model makes this adaptivity explicit and asks what threshold a rational CRN should choose *given* that the attacker will best-respond to it, rather than assuming a non-adaptive adversary.

Most existing PUEA games are posed as simultaneous-move (Nash) or Bayesian games between the SU/CRN and the attacker [5],[6],[9]. In many deployment scenarios, however, the CRN moves first in a structural sense: it publishes (or is constrained by regulation and hardware to use) a sensing threshold and access policy before individual attack events occur, and the attacker, observing the network's operating point over time, adapts its emulation power accordingly. This leader–follower structure is naturally captured by a **Stackelberg game** [16], and is the model we adopt here.

This paper extends the authors' original mathematical formulation of the PUEA energy-detection problem — the OFDM/AWGN signal model, the three-hypothesis sensing test ($H_0$: idle, $H_1$: PU present, $H_2$: PUEA present), and an initial Stackelberg utility pair for the CRN and the attacker — into a complete, closed-form equilibrium analysis, and couples that analysis with a reinforcement-learning mechanism for online adaptation. In doing so we identify and correct two structural issues in the naïve version of the game (Section 5) that, if left unaddressed, make the follower's problem unbounded and the leader's problem insensitive to the true cost of missed PU detection. The **contributions** of this paper are:

- A signal and hypothesis-testing model for energy-detection-based PUEA in OFDM-based CRNs, formalizing the false-alarm, detection, and attack-success probabilities under $H_0$, $H_1$, and $H_2$ (Section 3).
- Identification of an ill-posedness in the naïve Stackelberg formulation of PUEA (unbounded attacker power) and a principled fix via an exposure/power-cost term in the attacker's utility, together with a closed-form best-response power $P_m^*(\lambda)$ (Section 5.2).
- A previously unreported structural result: under the corrected follower problem, the attack success probability at equilibrium is invariant to the leader's threshold, for fixed attacker exposure cost (Section 5.3), which decouples the leader's threshold-selection problem from the direct effect of the attack.
- An augmented CRN utility that explicitly prices missed detection of the legitimate PU, and a closed-form, Neyman–Pearson-type expression for the equilibrium threshold $\lambda^*$ (Section 5.4).
- A Q-learning-based adaptive refinement of $\lambda^*$ that removes the requirement that the CRN know the attacker's channel gain and exposure cost exactly, together with the combined Stackelberg–Q-learning algorithm (Section 6).
- A numerical evaluation of the closed-form expressions across representative parameter ranges, used as an analytical consistency check pending full OFDM system-level and testbed validation (Section 8).

The remainder of the paper is organized as follows. Section 2 reviews related work. Section 3 presents the signal and sensing model. Section 4 formulates energy detection under PUEA. Section 5 develops the Stackelberg game and derives the closed-form equilibrium. Section 6 presents the Q-learning-based adaptive scheme.

Section 7 discusses computational complexity and convergence. Section 8 reports the numerical evaluation of the closed-form solution. Section 9 discusses implications, limitations, and threats to validity. Section 10 concludes and outlines the planned simulation and testbed campaign.

## RELATED WORK

### Detection- and Localization-Based Defenses

Early PUEA countermeasures focused on distinguishing a legitimate PU from an emulator using physical evidence unavailable to the attacker. Chen *et al.* proposed location-based verification using a distributed set of sensors together with a Neyman–Pearson test to bound the false-alarm rate under a benign-attacker assumption [2],[3]. Wireless localization and received-signal-strength (RSS)-based approaches similarly infer whether the transmitter is at the PU's known location; radio-frequency fingerprinting extends this idea to hardware-level waveform impairments that are costly for an attacker to spoof [10]. These approaches are effective against non-adaptive attackers but generally assume the attacker's strategy is fixed, which does not hold once the attacker can observe network behavior and adapt.

### Game-Theoretic Formulations

Jin, Anand and Subbalakshmi analyzed PUEA analytically and proposed hypothesis-testing-based detection without location infrastructure [4]. Tan, Sengupta and Subbalakshmi modeled the SU–attacker interaction as a non-cooperative dynamic multistage game and derived pure- and mixed-strategy Nash equilibria together with a belief-updating mechanism for the SU [5]. Li, Chakravarthy, Dehnie and Wu modeled PUEA as a stochastic game that destabilizes the queuing dynamics of the CRN and used Lyapunov drift as the stage reward, deriving explicit Nash strategies [6]. Li and Han's two-part "Dogfight in Spectrum" analysis treated PUEA as a zero-sum-like game under known and unknown channel statistics [7]. Nguyen-Thanh *et al.* and Ta *et al.* developed surveillance-game formulations in which the CRN allocates limited monitoring resources strategically across channels, deriving strong Stackelberg equilibrium policies for multichannel PUEA [8],[9]. Relative to this body of work, the present paper (i) works directly with the energy-detection statistics of an OFDM-based CRN rather than an abstract payoff table, (ii) explicitly identifies and corrects the unbounded-power pathology that arises when the follower's utility rewards attack success without a compensating cost, and (iii) derives closed-form equilibrium expressions rather than relying solely on numerical equilibrium search.

### Learning-Based and Hybrid Defenses

A separate line of work applies reinforcement learning (RL) to CR security, primarily for anti-jamming channel selection: Q-learning has been used to learn jammer-avoidance policies from wideband sensing feedback [12], and deep Q-network variants have been applied to jamming and spectrum-sensing-data-falsification (SSDF) attacks [11]. Surveys of machine-learning- and blockchain-based CR security more broadly report Q-learning-based policy optimization as improving

throughput relative to naïve fixed policies, and highlight PUEA-specific detection and prevention using RF fingerprinting combined with time–distance signal-strength evaluation. These works establish that RL is a viable tool for adapting sensing/access policy under unmodeled or slowly time-varying adversarial behavior, but, to the authors' knowledge, RL has not previously been combined with a closed-form Stackelberg equilibrium as a *warm start*, which is the combination proposed in Section 6 of this paper. More broadly, Manshaei *et al.*'s survey of game theory in network security and privacy [11] frames the leader–follower structure adopted here as appropriate whenever a defender must commit to an observable policy before an adaptive adversary responds — precisely the situation created by a publicly observable, hardware-constrained sensing threshold in a CRN.

## System and Signal Model

We consider a single-channel OFDM-based CRN with one licensed PU, one SU (or SU cluster reporting to a common fusion point), and one PUEA attacker. The SU senses the channel over a fixed sensing window and decides between three hypotheses:

$$H_0 : \text{channel idle}, \quad H_1 : \text{PU present}, \quad H_2 : \text{PUEA present (false occupancy)} \tag{1}$$

The baseband-equivalent received signal at the SU over the sensing window is

$$y(t) = h_p\, s_p(t) + h_m\, s_m(t) + n(t), \quad n(t) \sim (0, \sigma_n^2) \tag{2}$$

where $s_p(t)$ and $s_m(t)$ are the transmitted OFDM waveforms of the PU and the attacker respectively, $h_p$ and $h_m$ are the corresponding (assumed quasi-static, block-fading) channel gains at the SU, and $n(t)$ is additive white Gaussian noise of variance $\sigma_n^2$. Under $H_0$, $h_p s_p(t) = h_m s_m(t) = 0$; under $H_1$, $h_m s_m(t) = 0$; under $H_2$, both signal terms are present. Denote the transmit powers $P_p = [|s_p(t)|^2]$ and $P_m = \mathbb{E}[|s_m(t)|^2]$. Because the attacker mimics the PU's waveform structure, the SU cannot distinguish $s_p$ from $s_m$ on a per-symbol basis using energy statistics alone — the two signals are, by design of the attack, second-order indistinguishable to a plain energy detector.

## Energy Detection Under PUEA

The SU computes the sample energy of the received signal over $N$ samples,

$$E = (1/N) \sum_{n=1}^{N} |y(i)|^2 , \tag{3}$$

and compares it to a decision threshold $\lambda$: the channel is declared occupied if $E > \lambda$ and idle otherwise. Applying the central limit theorem to the sum of $N$ i.i.d. squared-magnitude samples (the standard large-$N$ Gaussian approximation used throughout the energy-detection literature), the conditional mean and variance of $E$ under each hypothesis are

$$[E \mid H_0] = \sigma_n^2, \quad \mathbb{E}[E \mid H_1] = |h_p|^2 P_p + \sigma_n^2, \quad \mathbb{E}[E \mid H_2] = |h_p|^2 P_p + |h_m|^2 P_m + \sigma_n^2 , \tag{4}$$

with common approximate standard deviation $\beta = \sigma_n\sqrt{(2/N)}$ under all three hypotheses (equal-variance Gaussian approximation). The resulting detection statistics are

$$Pf(\lambda) = Q( (\lambda - \sigma_n^2) / \beta ), \tag{5}$$

$$Pd(\lambda) = Q( (\lambda - |h_p|^2 P_p - \sigma_n^2) / \beta ), \tag{6}$$

$$Pa(\lambda, P_m) = Q( (\lambda - |h_p|^2 P_p - |h_m|^2 P_m - \sigma_n^2) / \beta ), \tag{7}$$

where $Q(x) = (1/\sqrt{2\pi}) \int_x^{\infty} e^{(-t^2/2)}\, dt$ is the Gaussian Q-function, Pf is the false-alarm probability ($H_0$ misclassified as occupied), Pd is the correct-detection probability of the genuine PU ($H_1$), and Pa is the probability that the SU is fooled into believing the channel is occupied by the PU when in fact it is the attacker ($H_2$) — i.e., the attack success probability. Equations (5)–(7) recover the sensing statistics used in the base derivation and match the structure of classical Neyman–Pearson energy detection [3],[4] with the attacker's contribution entering only through the mean shift $|h_m|^2 P_m$ in (7).

Two immediate observations motivate the game-theoretic treatment of Section 5. First, Pf and Pd move in the *same* direction as λ increases (both decrease), so the CRN faces the usual detector trade-off: raising λ reduces false alarms at the cost of missed detection of the genuine PU. Second, and specific to PUEA, Pa is *strictly increasing in* $P_m$ for any fixed λ — a rational attacker with no cost of transmission would therefore always benefit from increasing $P_m$ without bound, since $Q(\cdot) \to 1$ as its argument $\to -\infty$. A defense analysis that treats λ as the only free variable and $P_m$ as exogenous will therefore either underestimate the attacker's incentive to raise power, or (if it does optimize over $P_m$ under a bare success-probability objective) admit no finite optimal attack — a point we return to formally in Section 5.2.

## STACKELBERG GAME FOR THRESHOLD SELECTION AND ATTACK MITIGATION

### Game Structure

We model the interaction as a two-player Stackelberg game with complete information about the statistical model (5)–(7), in which the players' cost/exposure parameters, introduced below, may be imperfectly known — this imperfection is what motivates the learning layer of Section 6.

- **Leader — the CRN.** Chooses the sensing threshold λ (and, implicitly, the resulting access policy) to maximize a utility that rewards legitimate SU throughput and penalizes both false alarms and successful attacks.
- **Follower — the attacker.** Observes λ (or, more realistically, learns the CRN's effective operating point through repeated interaction) and chooses emulation power $P_m \geq 0$ to maximize attack effectiveness net of the cost of transmitting at high power.

The base utility pair, before correction, is

$$U_{CRN}(\lambda) = R_s(1 - Pf(\lambda)) - C_d\, Pf(\lambda) - C_m\, Pa(\lambda, P_m), \quad (8)$$

$$U_A(P_m) = Pa(\lambda, P_m) - C_a\, Pd(\lambda), \quad (9)$$

where $R_s$ is the reward for successful legitimate spectrum access, $C_d$ penalizes false alarms, $C_m$ prices an undetected attack, and $C_a$ is a nominal cost the attacker incurs when the genuine PU is correctly detected. Equations (8)–(9) are a natural starting point, but they have two structural weaknesses that we address before solving for equilibrium.

### Well-Posedness of the Follower's Problem

In (9), $P_d(\lambda)$ does not depend on $P_m$ (it is a function of the genuine PU's link only), so for a fixed λ the term $-C_a P_d(\lambda)$ is a constant with respect to the attacker's

decision variable. Consequently, maximizing (9) over $P_m$ is equivalent to simply maximizing $Pa(\lambda,P_m)$ over $P_m$, and since $Pa$ is monotonically increasing and unbounded above as $P_m \to \infty$ (its argument in (7) $\to -\infty$ and $Q(\cdot) \to 1$), the unconstrained follower's problem has **no finite maximizer**: the naïve model predicts an attacker that transmits with unbounded power, which is both physically unrealistic and analytically degenerate (no interior Stackelberg equilibrium exists).

We resolve this by recognizing that higher emulation power is not free to a real attacker: it increases the probability of exposure through mechanisms outside the energy-detection statistic itself — e.g., higher probability of triangulation/localization, higher probability of RF-fingerprint mismatch being flagged by a secondary check [10], increased power-amplifier and battery cost, and increased regulatory/legal risk. We capture this parsimoniously with a linear exposure-cost term $\kappa \cdot P_m$, $\kappa > 0$, added to the attacker's utility:

$$U_A(P_m) = Pa(\lambda,P_m) - \kappa P_m - C_a\, Pd(\lambda)\,. \tag{10}$$

This single additional term restores well-posedness, as shown next, and reduces to the naïve model in the limit $\kappa \to 0^+$ (in which case the equilibrium formally degenerates, consistent with the observation above).

## Closed-Form Attacker Best Response

Let $x(\lambda,P_m) = (\lambda - |h_p|^2 P_p - |h_m|^2 P_m - \sigma_n^2)/\beta$ denote the argument of the Q-function in (7), so that $Pa = Q(x)$. Writing $\phi(x) = (1/\sqrt{2\pi})e^{\wedge}(-x^2/2)$ for the standard normal density and using $dQ/dx = -\phi(x)$, the first-order condition for (10) with respect to $P_m$ (holding $\lambda$ fixed) is

$$\partial U_A/\partial P_m = \varphi(x) \cdot (|h_m|^2/\beta) - \kappa = 0 \;\Rightarrow\; \varphi(x^*) = \kappa\beta / |h_m|^2\,. \tag{11}$$

Because $\phi(x)$ is symmetric and decreasing in $|x|$, (11) has two roots $\pm|x|$; *the second-order condition* $\partial^2 U_A/\partial P_m^2 = -x\cdot\varphi(x)\cdot|h_m|^4/\beta^2$ *is negative (a maximum) only for* $x < 0$, so the relevant root is

$$x^* = -\sqrt{(\,2\ln(\,|h_m|^2 / (\kappa\beta\sqrt{2\pi})\,)\,)}, \tag{12}$$

which is real and an interior maximum exists **if and only if** $|h_m|^2 > \kappa\beta\sqrt{2\pi}$ — i.e., the attacker's channel to the SU must be strong enough, relative to its own marginal exposure cost, for emulation to be worthwhile at all. When this condition fails, the best response is the corner solution $P_m = 0$ *(no attack), which is itself a useful and easily checked deterrence criterion for network design: a CRN can be made structurally unattractive to emulate by ensuring* $\kappa$ *(through, e.g., mandatory RF-fingerprint checks that raise exposure risk) exceeds* $|h_m|^2/(\beta\sqrt{2\pi})$ *for all plausible attacker geometries. Substituting* $x$ back into the definition of $x(\lambda,P_m)$ gives the closed-form best response

$$P_m^*(\lambda) = \max\{\,0\,,\;[\,\lambda - |h_p|^2 P_p - \sigma_n^2 - \beta x^*\,]\,/\,|h_m|^2\,\}\,. \tag{13}$$

## An Equilibrium Invariance Property

A notable consequence of (11)–(12) is that $x$ *does not depend on* $\lambda$ *— it is determined entirely by* $\kappa$, $\beta$ *and* $|h_m|^2$. *Since the equilibrium attack success probability is* $Pa = Q(x)$, *this means that, whenever the interior solution applies, the attack success probability at equilibrium is invariant to the CRN's choice of threshold: a rational, cost-sensitive attacker always adjusts* $P_m$ *so as to keep its own success probability at the fixed*

*operating point Q(x)* dictated by its exposure cost, regardless of where the CRN sets λ. This can be confirmed directly by applying the envelope theorem to $V(\lambda) = \max_{P_m}\{Pa(\lambda,P_m) - \kappa P_m\}$: at the optimum, $\partial Pa/\partial P_m = \kappa$ (from (11)), and

$$dPa(\lambda,P_m^*(\lambda))/d\lambda = \partial Pa/\partial\lambda\,|_{P_m^*} + \kappa \cdot dP_m^*/d\lambda = -\kappa/|h_m|^2 + \kappa \cdot (1/|h_m|^2) = 0\,. \quad (14)$$

This is, to the authors' knowledge, a previously unreported structural feature of the PUEA Stackelberg game. It has an important practical implication: under this cost model, adjusting λ alone cannot be used to suppress the *success rate* of an already-adapting attacker — only to trade off Pf against Pd for the genuine PU. Suppressing Pa* requires raising the attacker's marginal exposure cost κ (e.g., via complementary detection layers such as RF fingerprinting or localization [10]), not merely retuning the energy-detection threshold. This gives a precise, quantitative justification for why detection-only defenses and threshold-only defenses are each, alone, structurally insufficient against a strategic attacker, and why hybrid defenses combining several detection modalities are recommended in the survey literature [10].

**Correcting the Leader's Utility and Solving for λ***

Equation (8) also has a gap: it penalizes false alarms and successful attacks but does not price the harm of *missing* the genuine PU (interference to the licensed incumbent), which is the quantity Pd is actually meant to protect against. We augment the CRN's utility with an explicit interference-cost term proportional to the missed-detection probability (1 – Pd(λ)):

$$U\_CRN(\lambda) = R_s(1 - Pf(\lambda)) - C\_d\,Pf(\lambda) - C\_m\,Pa(\lambda,P_m^*(\lambda)) - C\_i\,(1 - Pd(\lambda))\,. \quad (15)$$

By the invariance result of Section 5.4, $Pa(\lambda,P_m(\lambda)) = Q(x)$ is a constant with respect to λ (in the interior-solution regime), so it drops out of the leader's first-order condition and the threshold-selection problem reduces to a classical asymmetric-cost detector design problem:

$$\max_\lambda \{ R_s(1 - Pf(\lambda)) - C\_d\,Pf(\lambda) - C\_i\,(1 - Pd(\lambda)) \}\,. \quad (16)$$

Writing $x\_f = (\lambda-\sigma_n^2)/\beta$ and $x\_d = (\lambda-|h_p|^2P_p-\sigma_n^2)/\beta = x\_f - a/\beta$ with $a \triangleq |h_p|^2P_p$, differentiating (16) and setting the result to zero gives

$$(R_s+C\_d)\,\varphi(x\_f) = C\_i\,\varphi(x\_d)\,. \quad (17)$$

Using $x\_d = x\_f - a/\beta$ and expanding $\phi(x\_d)/\phi(x\_f) = \exp[-(x\_d^2-x\_f^2)/2] = \exp[\,x\_f\,a/\beta - a^2/2\beta^2\,]$, (17) can be solved in closed form for x_f, and hence for λ*:

$$\lambda^* = \sigma_n^2 + a/2 - (\beta^2/a)\cdot \ln(\,C\_i\,/\,(R_s+C\_d)\,)\,,\quad a = |h_p|^2P_p,\ \beta^2 = 2\sigma_n^2/N\,. \quad (18)$$

Equation (18) is the central analytical result of the leader's problem. It has an intuitive reading: the equilibrium threshold sits at the *midpoint* $\sigma_n^2 + a/2$ between the mean energy under $H_0$ and under $H_1$ — the classical equal-variance, equal-prior Bayes threshold — plus a correction term of order $\beta^2/a = O(1/N)$ that shifts λ *toward the noise floor when interference cost* $C_i$ *is high relative to* $(R_s+C\_d)$*, and toward the PU-present side when the reward/false-alarm cost dominates. Because the correction term scales as 1/N, it vanishes as the sensing window lengthens, so that* λ converges to the pure midpoint threshold as $N \to \infty$**, exactly as classical asymptotic detection theory predicts for a Gaussian equal-variance test. This convergence is used in Section 8 as an internal consistency check on the derivation.

**Equilibrium Summary**

Collecting (12), (13) and (18), the Stackelberg equilibrium ($\lambda$, $P_m$) of the corrected game satisfies

$$\lambda^* = \sigma_n^2 + a/2 - (\beta^2/a)\, ln(\, C\_i/(R_s+C\_d)\,), \quad x^* = -\sqrt{(\,2\, ln(\,|h_m|^2/(\kappa\beta\sqrt{2\pi})\,)\,)}, \quad (19)$$

$$P_m^* = max\{\, 0, [\,\lambda^* - a - \sigma_n^2 - \beta x^*\,] / |h_m|^2\,\}, \quad Pa^* = Q(x^*)\,, \quad (20)$$

and it is straightforward to verify that neither player can improve its utility by a unilateral deviation, i.e., U_CRN($\lambda$) $\geq$ *U_CRN($\lambda$) for all feasible $\lambda$ given* $P_m$($\lambda$), and U_A($P_m$) $\geq$ *U_A($P_m$) for all* $P_m \geq 0$ *given* $\lambda$, so that ($\lambda$,$P_m$) is indeed a Stackelberg equilibrium of the corrected game in the sense required by the original formulation.

**Q-LEARNING-BASED ADAPTIVE REFINEMENT**

**Motivation**

Equations (19)–(20) require the CRN to know $\sigma_n^2$, a = $|h_p|^2P_p$, N, and — critically — the attacker's channel gain $|h_m|^2$ and exposure-cost coefficient $\kappa$, neither of which is directly observable in practice. Channel gains drift with mobility and shadowing, and $\kappa$ is a property of the attacker's (unknown) risk model. We therefore treat (19)–(20) as an **analytical warm start** rather than a value to be used open-loop, and use a model-free Q-learning agent at the CRN to track the true, possibly time-varying, equilibrium threshold from observed sensing outcomes.

**Markov Decision Process Formulation**

We formulate threshold adaptation as a Markov Decision Process (MDP) ($\mathcal{S}$, $\mathcal{A}$, r, $\gamma$):

- **State** $s_t \in \mathcal{S}$: a quantized summary of recent sensing behavior at time step $t$, e.g., a tuple ($\hat{\lambda}_t$, $\hat{P}f_t$, $\hat{P}d_t$) formed from a sliding-window empirical estimate of the false-alarm and detection rates together with the current threshold bucket, discretized into a finite grid.
- **Action** $a_t \in \mathcal{A}$: a bounded threshold adjustment, $\mathcal{A}$ = {–$\Delta$, –$\Delta$/2, 0, +$\Delta$/2, +$\Delta$}, applied as $\lambda_{t+1} = \lambda_t + a_t$, so that the agent explores in the neighborhood of the current operating point rather than re-searching the full range at every step.
- **Reward** $r_t$: the empirical, per-step counterpart of (15), computed from observed spectrum-access outcomes and any available ground truth or higher-confidence secondary check (e.g., an occasional RF-fingerprint verification) used to label false alarms, missed detections and successful attacks over the window.
- **Discount factor** $\gamma \in (0,1)$: weights future reward, reflecting that the attacker may adapt over multiple interaction rounds rather than a single step.

The agent is initialized with $\lambda_0 = \lambda^*$ from (19) (the closed-form warm start) rather than an arbitrary or random threshold, which both reduces the exploration burden and guarantees that the online policy never performs worse than an uninformed baseline during the early, high-variance phase of learning.

**Q-Learning Update**

The agent maintains a tabular action-value function Q(s,a) and updates it after each observed transition ($s_t$, $a_t$, $r_t$, $s_{t+1}$) using the standard temporal-difference rule [14],[15]:

$$Q(s_t,a_t) \leftarrow Q(s_t,a_t) + \alpha\,[\, r_t + \gamma \max_{a'} Q(s_{t+1},a') - Q(s_t,a_t)\,], \quad (21)$$

with learning rate $\alpha \in (0,1]$, typically decayed over time (e.g., $\alpha_t = \alpha_0/(1+ct)$) to satisfy the Robbins–Monro conditions $\Sigma\alpha_t = \infty$, $\Sigma\alpha_t^2 < \infty$ that underlie standard Q-learning convergence guarantees for finite MDPs [14]. Action selection uses an $\varepsilon$-greedy policy with $\varepsilon$ annealed from an initial exploration rate toward a small floor value, so that the agent continues to probe the attacker's response at a low rate even after convergence, guarding against a subsequent shift in the attacker's $\kappa$ or $|h_m|^2$.

## Combined Algorithm

Algorithm 1 summarizes the complete Stackelberg-guided Q-learning (SGQL) procedure.

**Algorithm 1: Stackelberg-Guided Q-Learning (SGQL) for Adaptive PUEA Mitigation**

Input: $\sigma_n^2$, estimates of $a = |h_p|^2 P_p$, N, $R_s$, C_d, C_i, ($\hat{\kappa}$, $|\hat{h}_m|^2$ if available)
1: Compute warm-start threshold $\lambda_0 \leftarrow \lambda^*$ via Eq. (19)
2: Initialize $Q(s,a) \leftarrow 0$ for all (s,a); set $\varepsilon \leftarrow \varepsilon_0$, $\alpha \leftarrow \alpha_0$, $t \leftarrow 0$
3: loop (each sensing window)
4: Observe state $s_t$ from recent ($\hat{\lambda}$, $\hat{P}f$, $\hat{P}d$) statistics
5: With prob. $\varepsilon$ choose $a_t$ uniformly from $\mathcal{A}$; else $a_t \leftarrow \text{argmax\_a } Q(s_t,a)$
6: Apply $\lambda_{t+1} \leftarrow \text{clip}(\lambda_t + a_t, \lambda\_\text{min}, \lambda\_\text{max})$
7: Run energy detector at $\lambda_{t+1}$ for the next window; label outcomes
8: Compute reward $r_t$ via the empirical form of Eq. (15)
9: Observe next state $s_{t+1}$
10: Update $Q(s_t,a_t)$ via Eq. (21)
11: Decay $\varepsilon$, $\alpha$ ; $t \leftarrow t+1$
12: end loop
Output: adaptive threshold sequence $\{\lambda_t\}$, converging in probability to a neighborhood of the true (possibly time-varying) equilibrium $\lambda^*$

## Complexity and Convergence Discussion

**Closed-form component.** Evaluating (19)–(20) requires only elementary arithmetic and a logarithm/square-root evaluation, i.e., O(1) per re-computation; it can be recomputed on every window at negligible cost whenever updated estimates of $\sigma_n^2$, a, $\hat{\kappa}$ and $|\hat{h}_m|^2$ are available (e.g., from periodic PU-absence noise-floor calibration and any available attacker-channel estimate).

**Learning component.** With a discretized state space of size $|\mathcal{S}|$ and action space $|\mathcal{A}| = 5$, the tabular Q-learning update (21) is O(1) per step and the table requires $O(|\mathcal{S}||\mathcal{A}|)$ memory, which is modest for a coarse grid (e.g., $|\mathcal{S}|$ on the order of $10^2$–$10^3$ bins). Standard results for tabular Q-learning guarantee convergence to the optimal action-value function with probability 1 under the stated step-size conditions and infinite visitation of every state–action pair [14],[15]; in a non-stationary environment (a truly adaptive attacker whose $\kappa$ or $|h_m|^2$ drifts over time) these guarantees only hold locally between drifts, which is precisely why the closed-form $\lambda^*$ is retained as a periodically recomputed anchor/warm-start rather than discarded once learning begins — it re-centers the agent's search region whenever fresh estimates of the physical parameters become available, bounding the regret incurred during re-adaptation.

**End-to-end overhead.** The combined scheme adds only a small table lookup/update and an occasional closed-form recomputation to the existing energy-detection pipeline, so its computational footprint is compatible with real-time operation on typical CRN sensing hardware; the dominant cost, as in any energy-detection system, remains the N-sample energy computation (3) itself, which is unchanged by the proposed defense.

**Numerical Evaluation of the Closed-Form Solution**

Because the reinforcement-learning component and full OFDM physical-layer validation are the subject of an ongoing simulation and testbed campaign (Section 10), this section restricts itself to a **numerical evaluation of the closed-form analytical expressions derived in Section 5** — i.e., direct evaluation of Eqs. (12), (13), (18)–(20) over representative parameter ranges. The purpose is twofold: (i) to verify that the closed-form solution behaves as the derivation predicts (monotonicity, boundary/corner-solution behavior, asymptotic convergence), and (ii) to give the reader concrete intuition for the magnitude of the equilibrium quantities before Monte Carlo and hardware validation are performed. These are *not* Monte Carlo simulation results and no OFDM waveform, channel, or attacker behavior was simulated to produce them; they are exact evaluations of the closed-form formulas above and are reported as such.

### Leader's Threshold: Sensitivity to Interference Cost and Sensing Window

Table 1 evaluates $\lambda$ *from (18) for* $\sigma_n^2 = 1$, $a = |h_p|^2 P_p = 1$ *(0 dB average PU SNR),* $R_s = 10$, $C\_d = 5$, *over a sweep of the interference-cost weight* $C^*_i$, for a short sensing window (N = 10 samples) and a long one (N = 1000 samples).

| $C_i$ | $\lambda^*$ (N = 10) | $Pf(\lambda^*)$ | $Pd(\lambda^*)$ | $\lambda^*$ (N = 1000) |
|---|---|---|---|---|
| 1 | 2.0416 | 0.0099 | 0.4629 | 1.50542 |
| 5 | 1.7197 | 0.0538 | 0.7346 | 1.50220 |
| 10 | 1.5811 | 0.0969 | 0.8255 | 1.50081 |
| 20 | 1.4425 | 0.1612 | 0.8937 | 1.49942 |

*Table 1. Closed-form equilibrium threshold $\lambda^*$, and the resulting false-alarm and detection probabilities, as a function of the interference-cost weight $C_i$, for a short (N = 10) and long (N = 1000) sensing window. $\sigma_n^2 = 1$, $a = 1$, $R_s = 10$, $C\_d = 5$.*

Two predicted properties are confirmed. First, for the short window (N = 10), increasing $C_i$ — i.e., weighting harm to the licensed PU more heavily — monotonically lowers $\lambda$, *which correctly raises* $Pd$ *(better protection of the incumbent) at the cost of a higher* $Pf$ *(more spectrum opportunities lost to false alarms); this is the expected receiver-operating-characteristic trade-off and confirms that (18) is not, e.g., inverted in sign. Second, the N = 1000 column shows* $\lambda$ converging tightly around the pure midpoint value $\sigma_n^2 + a/2 = 1.5$ regardless of $C_i$, exactly as predicted by the $O(1/N)$ scaling of the cost-sensitivity term in (18); this reproduces the classical asymptotic

behavior of Bayes-optimal Gaussian thresholds and serves as an internal consistency check on the derivation rather than a claim about system performance.

**Follower's Best Response: Sensitivity to Exposure Cost**

Table 2 evaluates the attacker's closed-form best response (12)–(13) for a fixed leader threshold $\lambda = 1.72$ (the equilibrium value from the $C_i = 5$, $N = 10$ row of Table 1), $\sigma_n^2 = 1$, $a = 1$, $|h_m|^2 = 1$, over a sweep of the exposure-cost coefficient $\kappa$. The critical value above which no interior solution exists is $\kappa\_crit = 1/(\beta\sqrt{2\pi}) \approx 0.892$ for these parameters.

| $\kappa$ | $x^*$ | $P_m^*$ | $Pa^* = Q(x^*)$ |
|---|---|---|---|
| 0.1 | −2.0921 | 0.6556 | 0.9818 |
| 0.2 | −1.7293 | 0.4934 | 0.9581 |
| 0.3 | −1.4763 | 0.3802 | 0.9301 |
| 0.5 | −1.0760 | 0.2012 | 0.8590 |
| 0.7 | −0.6964 | 0.0314 | 0.7569 |

*Table 2. Closed-form attacker best response as a function of the exposure-cost coefficient $\kappa$, for $\lambda = 1.72$, $\sigma_n^2 = 1$, $a = 1$, $|h_m|^2 = 1$ (interior-solution regime, $\kappa < 0.892$).*

As predicted, both the equilibrium emulation power $P_m$ *and the equilibrium attack success probability* $Pa$ decrease monotonically as $\kappa$ increases — a more exposure-averse or more heavily penalized attacker rationally attacks less aggressively and succeeds less often. $P_m$ *approaches its lower bound of zero as* $\kappa$ *approaches* $\kappa\_crit$*, consistent with the corner-solution behavior noted after (12): for* $\kappa \geq \kappa\_crit$ *the closed form (13) would return a negative power, which is physically inadmissible and is correctly interpreted, per Section 5.3, as the attacker's rational choice not to attack at all (*$P_m = 0$*).* This transition — from active emulation at low exposure cost to complete deterrence at high exposure cost — is precisely the quantitative deterrence criterion referenced in Section 5.4 and is, again, a property of the closed-form solution rather than a simulated outcome.

**Scope and Limitations of This Evaluation**

The evaluation above validates the internal mathematical consistency of the derivation (correct monotonicity, correct corner-solution behavior, correct asymptotic limit) but does **not** constitute evidence about detection performance under a real OFDM waveform, multipath/fading channel, imperfect synchronization, non-Gaussian interference, or an attacker that deviates from the assumed linear exposure-cost model — nor does it exercise the Q-learning component of Section 6, whose behavior depends on the state discretization, reward shaping, and non-stationarity of a real attacker and can only be meaningfully assessed through Monte Carlo simulation and, ultimately, over-the-air testbed measurement. That validation is described as the immediate next step of this research in Section 10 and is deliberately not anticipated here.

## DISCUSSION

### Implications for CRN Design

The invariance result of Section 5.4 has a direct engineering implication: operators who rely solely on retuning the energy-detection threshold in response to observed PUEA activity are, under the cost model adopted here, chasing a quantity (the equilibrium attack success probability) that a rational attacker will restore by adjusting its own transmit power. Effective long-run mitigation therefore requires raising the attacker's *marginal cost of emulation* — through complementary detection layers such as RF fingerprinting, localization, or cooperative multi-sensor fusion [2],[10] — rather than through threshold adjustment alone. The role of the threshold, once this is recognized, is correctly understood as managing the false-alarm/missed-detection trade-off for the *genuine* PU (Eq. (18)), while a separate exposure-raising mechanism is required to suppress the attacker's incentive.

### Role of the Learning Layer

The closed-form results of Section 5 are derived under a specific parametric cost model (linear exposure cost, Gaussian equal-variance energy statistics) and assume the CRN can estimate a, $\kappa$ and $|h_m|^2$. In practice the attacker's cost structure is unobservable and may itself vary across attack campaigns or attacker types (selfish vs. purely malicious, as distinguished in the PUEA taxonomy of [4]). The Q-learning layer of Section 6 is designed precisely to relax this dependence: it treats the closed-form equilibrium as a *structural prior* — informing the warm start and the shape of the reward — while allowing the realized threshold trajectory to depart from (19) as evidence accumulates that the true environment differs from the assumed model. This hybrid analytical/learning design is intended to combine the sample efficiency and interpretability of the closed-form solution with the model-robustness of a data-driven policy, in the spirit of prior hybrid ML/game-theoretic proposals for CR security [11],[12].

### Limitations and Threats to Validity

- **Modeling assumptions.** The Gaussian, equal-variance approximation to the energy-detection statistic is standard but becomes less accurate at very small N or under strongly non-Gaussian interference; the linear exposure-cost model for the attacker is a parsimonious modeling choice, not an empirically calibrated one.
- **Single-attacker, single-channel scope.** The present analysis treats one attacker and one channel; multichannel PUEA and coalitions of attackers, addressed elsewhere via surveillance games [8],[9], are natural extensions (Section 10).
- **No system-level or hardware validation yet.** As emphasized in Section 8.3, the numerical results reported here evaluate the closed-form expressions only; they do not yet demonstrate performance under a full OFDM PHY, fading channel, or over-the-air attacker.
- **Learning-layer assumptions.** Convergence guarantees for the Q-learning component rely on stationarity assumptions between parameter drifts that may not hold against a highly adaptive, learning-capable attacker; adversarial robustness of the learning layer itself is left as future work.

## CONCLUSION AND FUTURE WORK

This paper developed a closed-form Stackelberg game-theoretic framework for mitigating Primary User Emulation Attacks in OFDM-based cognitive radio networks. Starting from an energy-detection signal model and an initial leader/follower utility pair, we identified and corrected two structural gaps — an unbounded attacker objective and a leader utility that omitted the cost of missed PU detection — and derived closed-form expressions for the attacker's best-response emulation power, the equilibrium sensing threshold, and the equilibrium attack success probability. A previously unreported invariance property (equilibrium attack success probability is independent of the leader's threshold, for fixed attacker exposure cost) was established analytically and confirmed numerically, along with the expected convergence of the equilibrium threshold to the classical midpoint detector as the sensing window lengthens. We further proposed a Q-learning-based adaptive layer that uses the closed-form equilibrium as a warm start and continues to track the true operating point without requiring exact knowledge of the attacker's channel and cost parameters, and we analyzed the computational complexity and convergence properties of the combined scheme.

The immediate next stage of this research, already scoped at the time of writing, is empirical validation in three steps: (i) Monte Carlo simulation of the full OFDM PHY layer (realistic subcarrier structure, cyclic prefix, multipath/fading channel, and synchronization impairments) to test the Gaussian energy-detection approximation underlying (5)–(7) under realistic conditions; (ii) agent-based simulation of the Q-learning layer of Algorithm 1 against both a rational best-responding attacker (to validate convergence to the predicted equilibrium) and non-rational/adaptive attacker models (to stress-test robustness beyond the assumed cost structure); and (iii) a software-defined-radio testbed (e.g., USRP-based, in the spirit of prior PUEA testbed studies [13]) to validate detection performance, threshold adaptation latency, and false-alarm/detection trade-offs over the air. Beyond empirical validation, natural theoretical extensions include a multichannel formulation combining the present threshold game with the surveillance-resource-allocation games of [8],[9], a Bayesian (incomplete-information) treatment of the attacker's exposure-cost parameter $\kappa$, and a deep-Q or actor–critic extension of Section 6 to support continuous threshold and multichannel action spaces.

## REFERENCES

J. Mitola, “Cognitive radio for flexible mobile multimedia communications,” in Proc. IEEE Int. Workshop Mobile Multimedia Communications, 1999, pp. 3–10.

R. Chen, J.-M. Park, and J. H. Reed, “Defense against primary user emulation attacks in cognitive radio networks,” IEEE J. Sel. Areas Commun., vol. 26, no. 1, pp. 25–37, Jan. 2008.

R. Chen and J.-M. Park, “Ensuring trustworthy spectrum sensing in cognitive radio networks,” in Proc. IEEE Workshop Networking Technologies for Software Defined Radio Networks, Sept. 2006, pp. 110–119.

Z. Jin, S. Anand, and K. P. Subbalakshmi, "Detecting primary user emulation attacks in dynamic spectrum access networks," in Proc. IEEE Int. Conf. Communications (ICC), 2009.

Y. Tan, S. Sengupta, and K. P. Subbalakshmi, "Primary user emulation attack in dynamic spectrum access networks: A game-theoretic approach," IET Commun., vol. 6, no. 8, pp. 964–973, 2012.

H. Li, V. Chakravarthy, S. Dehnie, and Z. Wu, "Primary user emulation attack game in cognitive radio networks: Queuing aware dogfight in spectrum," in Game Theory for Networks (GameNets 2012), LNICST vol. 105, Springer, 2012, pp. 1–11.

H. Li and Z. Han, "Dogfight in spectrum: Combating primary user emulation attacks in cognitive radio systems, Part I: Known channel statistics," IEEE Trans. Wireless Commun., vol. 9, no. 11, pp. 3566–3577, 2010.

N. Nguyen-Thanh, P. Ciblat, A. T. Pham, and V. T. Nguyen, "Surveillance strategies against primary user emulation attack in cognitive radio networks," IEEE Trans. Wireless Commun., vol. 14, no. 9, pp. 4981–4993, 2015.

D. T. Ta, N. Nguyen-Thanh, P. Maillé, P. Ciblat, and V. T. Nguyen, "Strategic surveillance against primary user emulation attacks in cognitive radio networks," IEEE Trans. Cogn. Commun. Netw., vol. 4, no. 3, pp. 582–596, 2018.

S. U. Rehman, K. W. Sowerby, and C. Coghill, "Radio-frequency fingerprinting for mitigating primary user emulation attack in low-end cognitive radios," IET Commun., vol. 8, no. 8, pp. 1274–1284, 2014.

M. H. Manshaei, Q. Zhu, T. Alpcan, T. Başar, and J.-P. Hubaux, "Game theory meets network security and privacy," ACM Comput. Surv., vol. 45, no. 3, pp. 1–39, 2013.

M. Ling, K.-L. Yau, J. Qadir, G. S. Poh, and Q. Ni, "Application of reinforcement learning for security enhancement in cognitive radio networks," Appl. Soft Comput., vol. 37, pp. 809–829, 2015.

E. Cadena Muñoz, E. Rodriguez-Colina, L. F. Pedraza, and I. P. Paez, "Detection of dynamic location primary user emulation on mobile cognitive radio networks using USRP," EURASIP J. Wireless Commun. Netw., vol. 2020, Art. 62, 2020.

R. S. Sutton and A. G. Barto, Reinforcement Learning: An Introduction, 2nd ed. Cambridge, MA: MIT Press, 2018.

C. J. C. H. Watkins and P. Dayan, "Q-learning," Machine Learning, vol. 8, pp. 279–292, 1992.

T. Başar and G. J. Olsder, Dynamic Noncooperative Game Theory, 2nd ed. Philadelphia, PA: SIAM, 1999.

Z. Chen, T. Cooklev, C. Chen, and C. Pomalaza-Raez, "Modeling primary user emulation attacks and defenses in cognitive radio networks," in Proc. IEEE Int. Performance Computing and Communications Conf. (IPCCC), 2009, pp. 208–215.