Moving Target Defense in Credit Card Fraud Detection: A Flat-Maximum Perspective

Abstract

Credit card fraud detection can be viewed as an attacker–defender game in which adaptive adversaries exploit the static nature of machine learning (ML) models. RESONANT, a deep reinforcement learning (DRL)-based moving target defense (MTD) framework, addresses this challenge by dynamically switching among multiple classifiers to confound adversaries. While such switching increases unpredictability for the attacker, it can incur a cost: deviating from the single, static model that is optimal under current conditions. Understanding and bounding this cost is critical for evaluating the practical viability of MTD strategies in high-volume financial transaction environments.

The flat-maximum effect—originally established in credit scoring research (Overstreet et al., 1992)—provides a theoretical basis for bounding the loss from sub-optimal switching. This effect, observed when performance surfaces are relatively flat near the optimum, implies that multiple decision boundaries can achieve near-optimal performance despite differing parameterizations. Recent research (Guerrero & Flores, 2023; Crook et al., 2022) confirms the persistence of flat-maximum-like behavior in modern scoring systems, including regularized logistic regression and certain ensemble methods, even under population drift. In the adversarial fraud context, this suggests that an MTD strategy can alter decision boundaries—thereby disrupting adversary learning—without materially sacrificing detection accuracy.

We extend RESONANT by explicitly quantifying the trade-off between resilience gain from boundary switching and performance loss due to sub-optimality. Using an 80-million-transaction real-world dataset, we simulate iterative attacker–defender interactions under varying fraud rates, label delay windows, and adversary adaptation speeds. Our results show that: (1) detection performance across multiple well-tuned classifiers lies within a narrow performance band, confirming a flat-maximum region; (2) DRL-based switching can exploit this region to introduce unpredictability while maintaining bounded loss—often <2% relative AUC degradation; and (3) under harsher attack conditions, resilience gains outweigh the marginal performance penalty, particularly when the adversary's learning horizon is long.

By framing boundary-switching costs within the flat-maximum effect, we provide a principled argument for MTD strategies in financial fraud defense. This connection bridges historical insights from credit scoring robustness with modern adversarial ML, offering both a conceptual justification and empirical evidence that resilience-oriented switching can be achieved without compromising operational effectiveness. The framework has broader applicability to other dynamic classification environments, where adaptive threats and model stability must be balanced in real time.

Keywords: Credit Card Fraud Detection, Moving Target Defense, Flat-Maximum Effect, Reinforcement Learning, Adversarial Machine Learning, Model Robustness, Consumer Finance