ONLINE LEARNING IN STACKELBERG SECURITY GAMES WITH TIME-VARYING ATTACK INTENSITIES

Anonymous authors
Paper under double-blind review

ABSTRACT

This work investigates no-regret online learning for Repeated Stackelberg Security Games (RSSGs) against attackers with time-varying attack intensities. Standard RSSG models often assume fixed attack intensities, limiting their ability to capture evolving attack capabilities. To bridge this gap, we formalize a model in which follower numbers, types, and attack intensities can vary across rounds. We derive an exact mixed-integer linear programming oracle under optimistic tie-breaking. Under full-information feedback, combining this oracle with online learning algorithm yields expected $\mathcal{O}(\sqrt{T})$ regret against adversarial attacker sequences. Under bandit feedback, where only aggregate attack is observed, we represent the payoffs of all candidate defense strategies through a small set of basis strategies. This representation enables unbiased estimation of all candidate strategies' payoffs from a single aggregate payoff per round. Building on this representation, we combine barycentric spanner with EXP2 to achieve expected $\mathcal{O}(\sqrt{T})$ regret against adversarial sequences. This guarantee contains no additional logarithmic factors in T and matches an $\Omega(\sqrt{T})$ lower bound established within our model. Numerical simulations demonstrate the effectiveness of the proposed algorithms.

1 INTRODUCTION

In the Stackelberg Security Game formulation (Conitzer & Sandholm, 2006; Brückner & Scheffer, 2011; Li & Sethi, 2017; Castiglioni et al., 2020), a leader (defender) first commits to a randomized patrolling or protection strategy, and a follower (attacker) observes and best responds to the leader's mixed strategy, by choosing a strategy to attack targets, in order to maximize its own payoff. This leader-follower paradigm has seen remarkable success in a wide range of domains, for instance, it has been deployed for surveillance patrol scheduling at Los Angeles International Airport (Pita et al., 2008), protection of critical infrastructures (Shieh et al., 2012), and wildlife conservation against illegal poaching (Yang et al., 2014). Beyond physical security, the same modeling principles have been extended to cyber-security challenges, including network resource allocation (Bai et al., 2023; Gan et al., 2022) and signaling scheme design (Castiglioni et al., 2020; Bernasconi et al., 2023).

However, existing studies and classical models typically assume that the attacker has a fixed attack capacity and a single target, which rarely holds in practice. In cyber attacks, for instance, the adversary can simultaneously attack or disable multiple servers, and this number of servers under attacks often grows over time as the adversary accumulates more compromised hosts (Antonakakis et al., 2017); Also, a network intruder who initially can only probe a single subnet may eventually launch a coordinated attack against multiple critical nodes (Ho et al., 2021). Analogously, in physical settings, poachers can maintain many snares simultaneously, and the scale of their operations may vary across seasons as hunting effort and resource availability change (MacMillan & Nguyen, 2014; Becker et al., 2013). In these cases, the attacker's capacity to strike multiple targets varies dynamically, driven by exogenous shifts or adaptive resource allocation. Such time-varying attack intensities break the standard assumption of a fixed attack capacity, and, in turn, invalidate the assumption of fixed follower feasible action set that classical regret-minimization frameworks rely on. Motivated by these real-world phenomena, we aim to extend the existing models to Stackelberg Security Games with attackers exhibiting time-varying attack intensities, and to develop methodologies that allow the defender to adaptively learn and minimize regret against dynamically evolving attacks.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

An additional challenge is to achieve optimal regret under partial-information feedback, where the defender must learn from aggregate attack outcomes without observing individual follower types or attack intensities. Personnat et al. (2025) establish an $\Omega(\sqrt{T})$ regret lower bound in a related Bayesian multi-follower Stackelberg setting. Meanwhile, the adversarial bandit guarantees of Bernasconi et al. (2023) and Balcan et al. (2025) in their respective models scale as $O(\sqrt{T \log T})$ and $O(\sqrt{T} \log T)$ respectively. These published bounds retain additional logarithmic factors in the horizon. This motivates the pursuit of an $O(\sqrt{T})$ regret guarantee while accommodating multiple followers and time-varying attack intensities. We achieve this guarantee under aggregate bandit feedback against adversarial follower sequences and establish a matching $\Omega(\sqrt{T})$ lower bound within our model. Consequently, our regret bound is minimax optimal in its dependence on T, without additional horizon dependent logarithmic factors.

Our Contributions.

We generalize the classical SSG model by allowing the attacker's attack capacity to vary over time. The contributions of this work are summarized as follows:

Exact MILP formulation for intensity-constrained best responses. We model the follower's best-response dynamics by introducing auxiliary decision variables and constraints to encode the attacker's best-response selection under intensity-constrained action capacities. Based on this model, we develop an MILP via Big-M linearization. This yields an exact and tractable optimization oracle for computing optimal defender strategies under dynamically varying attack intensities.

Online learning with adversarially evolving attackers. We then extend this model to a general online setting, where the number of attackers, types, and attack intensities all vary arbitrarily and adversarially across rounds. By integrating the MILP oracle with FPL, we obtain an efficient online algorithm that achieves sublinear regret despite the non-convex and discontinuous utility structure.

Optimal learning from aggregate attack feedback. When multiple followers act simultaneously, aggregate observations generally do not reveal their individual types or attack intensities. We show that the defender can still achieve optimal square-root regret by learning the payoffs of defense strategies directly. To do so, we express all candidate payoffs through a small set of basis strategies. This representation allows a single aggregate payoff to provide unbiased estimates for every candidate strategy, without recovering the underlying follower composition. We combine these estimates with EXP2 and barycentric spanner based exploration to handle adversarial changes in follower numbers, types, and attack intensities. Given game parameters, the resulting expected regret is $O(\sqrt{T})$, with no additional logarithmic factors in T. We also establish an $\Omega(\sqrt{T})$ lower bound within the same model, showing that this dependence on the horizon cannot be improved in general.

1.1 Related Works

Online learning in Stackelberg games. A growing number of work studies online learning in Stackelberg games, where early works (Letchford et al., 2009; Bacchiocchi et al., 2024; Blum et al., 2014; Zhang et al., 2022) focus on computing optimal defender strategies under repeated interactions. Some research has extended this framework to several dimensions; for instance, (Donahue et al., 2024) considers learning in cooperative Stackelberg settings, while (Harris et al., 2024; Balcan et al., 2025) incorporate side information and develop bandit learning algorithms. In adversarial settings, (Balcan et al., 2015; Harris et al., 2024; Balcan et al., 2025) investigate regret minimization against arbitrary sequences of attacker types under both full and partial feedback. However, their models are limited to a single follower per round with a fixed, single-action attack capacity, and fail to accommodate settings with multiple attackers or time-varying attack intensities. More recent works like (Personnat et al., 2025; Cheng et al., 2022) consider settings with multiple simultaneous followers. Nevertheless, they assume stochastic or distributional models over attackers, rather than considering the fully adversarial attacking sequences, and do not capture time-varying attack intensities.

Modeling diverse attack behaviors. To better capture attacker behaviors, several studies consider richer attacker capabilities through Bayesian models and application specific formulations. For example, Wahab et al. (2019) study heterogeneous attack types in a Bayesian framework, while Yu et al. (2025); Xu et al. (2026) investigate resource-constrained attacks in control and cyber-physical systems. Application specific models, such as DDoS defense in IoT networks (Chen et al., 2021), further approximate attack intensities via discretization or known distributions. Despite these

Table 1: Comparison of our results with related works under different settings.

Results	Feedback	Multiple followers	Multiple types	Dynamic l	Adversarial	Regret
Our Algorithms	Full Info	✓	✓	✓	✓	$\mathcal{O}(\sqrt{T})$
	Partial Info	✓	✓	✓	✓	$\mathcal{O}(\sqrt{T})$
(Balcan et al., 2015)	Full Info	✗	✓	✗	✓	$\mathcal{O}(\sqrt{T})$
	Partial Info	✗	✓	✗	✓	$\mathcal{O}(T^{2/3})$
(Bernasconi et al., 2023)	Partial Info	✓	✓	✗	✓	$\mathcal{O}(\sqrt{T\log T})$
(Harris et al., 2024)	Full Info	✗	✓	✗	✓	$\mathcal{O}(\sqrt{T\log T})$
	Partial Info	✗	✓	✗	✓	$\mathcal{O}(T^{2/3}(\log T)^{1/3})$
(Personnat et al., 2025)	Full Info	✓	✓	✗	✗	$\mathcal{O}(\sqrt{T})$
	Partial Info	✓	✓	✗	✗	$\mathcal{O}(\sqrt{T\log T})$
(Balcan et al., 2025)	Partial Info	✗	✓	✗	✓	$\mathcal{O}(\sqrt{T}\log T)$

advances, existing approaches remain limited in generality. Most works rely on strong structural assumptions, such as stochastic attack models or application-specific dynamics, which do not consider dynamic and adversarial variations in attacker capabilities.

Regret guarantees under bandit feedback. The adversarial bandit methods of Balcan et al. (2015); Harris et al. (2024) conduct exploration rounds within each time block and achieve $T^{2/3}$ dependence. Balancing the repeated exploration cost against regret from block-level updates yields this dependence on T. For contextual Stackelberg games, Balcan et al. (2025) obtain $O(\sqrt{T}\log T)$ regret through logdet-FTRL, whose regularization and covariance-estimation analysis retain logarithmic factors. In addition, a direct instantiation of the logdet-FTRL approach of Balcan et al. (2025) in our setting can be infeasible when structural dependencies make all feasible strategy matrices singular, leaving the log-determinant regularizer without a finite value. For multiple simultaneous followers, Personnat et al. (2025) obtain $O(\sqrt{T\log T})$ regret under partial information and establish a $\Omega(\sqrt{T})$ lower bound. In their UCB-based method, the additional $\sqrt{\log T}$ factor arises from confidence bounds used to control utility estimation errors. Within our model, we achieve $O(\sqrt{T})$ expected regret under bandit feedback against adversarial follower sequences, while allowing multiple simultaneous followers and time-varying attack intensities. This bound matches the $\Omega(\sqrt{T})$ lower bound established within our model (Proposition 2) and is therefore optimal in its dependence on T.

Table 1 summarizes these regret guarantees alongside the corresponding modeling assumptions and feedback settings, where l denotes the attack intensity.

2 RSSGs WITH TIME-VARYING ATTACK INTENSITY

SSG Modeling of Time-varying Attack Intensity. Under the time-varying attack intensity, the problem is formulated as an extended Repeated Stackelberg Security Game (RSSG). The interaction between the defender and the attacker is characterized as follows: (1) each edge corresponds to a target that may be attacked by the attacker, and (2) the weight allocation vector represents the defender's mixed strategy over targets. Based on this formulation, the local decision-making process of defender is modeled as an RSSG, denoted by $\mathcal{G}$.

Definition 1. *An RSSG is defined as a tuple $\mathcal{G} = (T, \mathcal{E}^-, \mathcal{E}_k, \mathcal{W}, \mathcal{U})$, where 1) T is the time horizon; 2) $\mathcal{E}^- = \{1, ..., N\}$ is set of targets with cardinality N; 3) $\mathcal{E}_k \subseteq \mathcal{E}^-$ denotes the subset of attacked targets selected by the attacker of type k; 4) $\mathcal{W} = \{\mathbf{w} \in \mathbb{R}^N_{\geq 0} \mid \sum_j w_j = 1\}$ is the defender's mixed strategy space where $\mathbf{w}$ is a weight allocation vector; and 5) $\mathcal{U}$ specifies the utility structure.*

The utility structure is defined as follows. If target j is attacked by attacker of type $k \in \{1, \ldots, K\}$ and not covered by the defender, the defender and the attacker receive utilities $\lambda_j^r \in [-1, 0]$ and $\lambda_j^k \in [0, 1]$, respectively. If target j is attacked and covered by the defender, the defender and the attacker receive utilities $\rho_j^r \in [0, 1]$ and $\rho_j^k \in [-1, 0]$, respectively. Given that the defender

protects target j with probability w_j, the expected utility of the type k attacker when attacking target j is $u_a^k(\mathcal{E}_j, \mathbf{w}) = \lambda_j^k(1 - w_j) + \rho_j^k w_j$. The attacker's maximum capacity under fully available resources is F, i.e., the attacker can choose to attack at most F targets at each time step. The attack intensity l, defined as the number of targets the attacker chooses to attack, is determined at the beginning of each round. Given the defender's mixed strategy $\mathbf{w}$, the attacker selects l targets to attack. The attacker's best response is defined as the subset of targets that maximizes its total utility: $\mathbf{b}_k(\mathbf{w}) = \arg\max_{\mathcal{E}_k \cap \mathcal{E}^-} \sum_{\mathcal{E}_j \in \mathcal{E}_k \cap \mathcal{E}^-} u_a^k(\mathcal{E}_j, \mathbf{w})$, s.t. $|\mathcal{E}_k \cap \mathcal{E}^-| = l$.

Throughout the paper, we use an optimistic tie-breaking rule and any remaining ties are broken by increasing target index. The defender and the attackers know each other's utility parameters. At the end of the round, full-information feedback reveals the realized follower types and intensities, whereas bandit feedback reveals only the aggregate attack outcome specified in Section 5.

3 Exact Computation for Extended RSSG

RSSG with time-varying attack intensity poses a new computational challenge: classic solvers like DOBSS Paruchuri et al. (2008) are no longer suitable because the follower action set changes with the realized intensity. In this section, we design an exact algorithm for the defender such that, given a prior distribution $\mathbf{q} = [q_1, \ldots, q_F]$ over attack intensities (e.g., the empirical frequencies), it computes the optimal strategy, where q_l represents the defender's prior probability for attacker of attack intensity $l \in \mathcal{L} = \{1, \ldots, F\}$. The current intensity is fixed before the defender makes its decision and is not observable at the time of decision making. We first formulate the problem as a bi-level optimization problem and then reduce it to an equivalent single-level MILP.

Bi-Level Formulation. We first consider a fixed attacker type k. For each attack intensity l, we introduce a binary decision vector to represent the attacker's best response, $\mathbf{h}_l^k = [h_{1,l}^k, \ldots, h_{N,l}^k]^\top \in \{0,1\}^N$, where $h_{j,l}^k = 1$ indicates that target $\mathcal{E}_j$ is selected by the attacker, and $h_{j,l}^k = 0$ otherwise. Let $f_l^k(\mathbf{w})$ denote the defender's expected utility against an attacker of type k and intensity l under the tie-breaking rule in Section 2, written as:

$$f_l^k(\mathbf{w}) = \sum_{\mathcal{E}_j \in \mathcal{E}^-} h_{j,l}^k \left[\lambda_j^r(1 - w_j) + \rho_j^r w_j\right]. \tag{3.1}$$

Given a prior attack intensity distribution $\mathbf{q}$, the defender aims to maximize its expected utility $\sum_{l=1}^F q_l f_l^k(\mathbf{w})$ by selecting an optimal strategy $\mathbf{w}$. This leads to the following bi-level formulation:

$$\max_{\mathbf{w} \in \mathcal{W},\, \mathbf{h}_{1:F}^k} \quad \sum_{l=1}^F q_l \sum_{\mathcal{E}_j \in \mathcal{E}^-} h_{j,l}^k \left[\lambda_j^r(1 - w_j) + \rho_j^r w_j\right] \tag{3.2a}$$

$$\text{s.t.} \quad \mathbf{h}_l^k \in \arg\max_{\substack{\mathbf{h}' \in \{0,1\}^N \\ \sum_j h_j' = l}} \sum_{\mathcal{E}_j \in \mathcal{E}^-} h_j' u_a^k(\mathcal{E}_j, \mathbf{w}), \quad l = 1, \ldots, F. \tag{3.2b}$$

In this formulation, (3.2a) represents the defender's upper-level objective. The constraint $\mathbf{w} \in \mathcal{W}$ ensures that the strategy vector remains within the probability simplex. The lower-level constraint (3.2b) captures the attacker's best response. Given a fixed strategy $\mathbf{w}$ and attack intensity l, the attacker selects l targets with the highest utilities. By introducing the response variables $\{\mathbf{h}_l^k\}_{l=1}^F$, the formulation unifies all possible attack intensities within a single optimization framework.

Single-Level MILP Reduction To enable efficient computation, we transform the bi-level problem into an equivalent single-level MILP. This is achieved by 1) linearizing the bilinear terms, and 2) encoding the attacker's best-response conditions using Big-M constraints, which is formulated as

$$\max_{\mathbf{w},\, \mathbf{z}_{1:F}^k,\, \mathbf{h}_{1:F}^k,\, \tau_{1:F}^k} \quad \sum_{l=1}^F q_l \sum_{\mathcal{E}_j \in \mathcal{E}^-} \left[\lambda_j^r h_{j,l}^k + (\rho_j^r - \lambda_j^r) z_{j,l}^k\right] \tag{3.3a}$$

$$\text{s.t.} \quad \mathbf{w} \in \mathcal{W}, \tag{3.3b}$$

$$z^k_{j,l} \le w_j, \quad z^k_{j,l} \ge w_j + h^k_{j,l} - 1, \tag{3.3c}$$

$$\sum_{\mathcal{E}_j \in \mathcal{E}^-} h^k_{j,l} = l, \quad \mathbf{h}^k_l \ge \mathbf{h}^k_{l-1} \quad (l \ge 2), \tag{3.3d}$$

$$u^k_a(\mathcal{E}_j, \mathbf{w}) \ge \tau^k_l - B(1 - h^k_{j,l}), \quad u^k_a(\mathcal{E}_j, \mathbf{w}) \le \tau^k_l + B h^k_{j,l}, \tag{3.3e}$$

$$z^k_{j,l} \in [0,1], \quad h^k_{j,l} \in \{0,1\}, \quad z^k_{j,l} \le h^k_{j,l}, \quad -1 \le \tau^k_l \le 1. \tag{3.3f}$$

The components and constraints of (3.3) are detailed as follows:

- *Linearized Payoff Structure* (3.3c) *and* (3.3f)*:* To address the nonlinearity introduced by the bilinear term $h^k_{j,l} w_j$, we introduce an auxiliary variable $z^k_{j,l}$ to represent their product. This enables a standard linearization via McCormick-type constraints, ensuring $z^k_{j,l} = h^k_{j,l} w_j$ at optimality. As a result, the payoff can be expressed as a linear combination of the reward term ρ^r_j and the penalty term λ^r_j.
- *Attacker's Cardinality and Nested Logic* (3.3d)*:* The cardinality constraint selects exactly l targets, while the nesting constraints impose a nested structure across attack intensities. For a fixed defender strategy and attacker type k, higher attack intensities expand upon previously selected targets.
- *Optimality Conditions via Big-M* (3.3e)–(3.3f)*:* These conditions represent the optimality conditions for the attacker's best response (3.2b). For each l, a scalar variable τ^k_l is introduced to represent a utility threshold: any target with attacker utility $\lambda^k_j(1 - w_j) + \rho^k_j w_j > \tau^k_l$ is forced to be selected ($h^k_{j,l} = 1$), while targets with utility below the threshold are set to $h^k_{j,l} = 0$. Equality at the threshold is permitted, allowing tie-breaking in favor of defender. Here, B is a sufficiently large positive constant.

4 Online Learning with Full Information Feedback

Based on the above optimization framework, we consider a more general and challenging setting in which the attacker sequence evolves arbitrarily over time. In this case, the number of attackers, their types, and their attack intensities vary over time, even potentially in an adversarial manner. Existing works only partially address this setting. Prior studies on adversarial Stackelberg learning and multiple-attacker models (e.g., Balcan et al. (2015); Harris et al. (2024); Personnat et al. (2025)) do not consider time-varying attack intensities, while works incorporating attack intensities (e.g., Qian et al. (2016); Xu et al. (2026)) rely on structured dynamics and cannot handle adversarially chosen attacker sequences or multiple followers. This gap motivates us to study a unified setting that simultaneously captures (i) multiple attackers with diverse attacker types, (ii) time-varying attack intensities, and (iii) adversarially evolving attacker sequences.

Attackers with Arbitrary Types and Intensities. We consider a setting where at each time step t, the number of followers $n(t) \le C$, their types, and their attack intensities can be chosen arbitrarily. The follower configuration is non-anticipating: it may depend on past interactions but is fixed before the defender uses its current private randomness to choose $\mathbf{w}(t)$. This setting captures practical situations in which attacker characteristics, including their number, types, and attack intensities, change over time due to resource constraints, coordination effects, and environmental uncertainty.

Under full-information feedback, the current follower configuration remains hidden while the defender chooses $\mathbf{w}(t)$. After observing this strategy, the followers select their attack targets according to the prescribed best responses. The configuration is revealed only at the end of the round, when the defender observes every follower's type and attack intensity and updates the empirical count matrix.

Let $\mathbf{M} \in \mathbb{R}^{K \times F}$ be the matrix with entrie $(\mathbf{M})_{k,l} = m_{kl}$, where m_{kl} denotes the number of followers of type k with intensity l. Similarly, define $\mathbf{F}(\mathbf{w}) \in \mathbb{R}^{K \times F}$ with entries $(\mathbf{F}(\mathbf{w}))_{k,l} = f^k_l(\mathbf{w})$. The total utility can then be compactly expressed as the sum of element-wise products $\sum_{k=1}^{K} \sum_{l=1}^{F} \big(\mathbf{M} \odot \mathbf{F}(\mathbf{w})\big)_{kl}$, where $\odot$ denotes the Hadamard (entrywise) product.

At round t, let $\mathbf{M}(t)$ denote the realized follower count matrix and define its empirical average as $\hat{\mathbf{M}}(t) = \frac{1}{t} \sum_{t'=1}^{t} \mathbf{M}(t')$. Their entries are denoted by $m_{kl}(t)$ and $\hat{m}_{kl}(t)$, respectively. Empirical average utility of strategy $\mathbf{w}$ is obtained by replacing $\mathbf{M}$ with $\hat{\mathbf{M}}(t)$ in the above utility expression.

Exact Computation for Attacker Sequences. Given the empirical count matrix $\hat{\mathbf{M}}(t)$, we extend the single-level MILP in (3.3) to multiple attacker types. Specifically, we impose the linearization and best-response constraints for each type while retaining a common defender strategy $\mathbf{w}$. We weight each payoff $f_l^k(\mathbf{w})$ by empirical count $\hat{m}_{kl}(t)$. The resulting oracle computes an exact maximizer of the defender's empirical average utility. The complete MILP formulation is provided in Appendix F.

Proposition 1 (Oracle size and enumeration bound)**.** *The MILP oracle F has $O(KFN)$ variables and constraints. It admits an exact enumeration-based implementation that solves at most $\big(N!/(N-F)!\big)^K = O(N^{KF})$ linear programs, each with $O(KFN)$ variables and constraints.*

This bound applies to the enumeration-based LP implementation. The compact MILP formulation allows direct solution using standard MILP solvers such as Gurobi (Gurobi Optimization, LLC, 2024). These solvers can exploit presolve, cutting planes, and branch-and-bound to prune the search space without requiring explicit enumeration of all follower response profiles, yielding substantial computational savings in practice. Our runtime experiments in Appendix G further support this observation, showing that the compact MILP formulation achieves substantially shorter solution times than LP enumeration on the larger tested instances.

However, deterministic strategy selection based on past empirical counts can make the defender's next allocation predictable, and adversaries may exploit this predictability by choosing unfavorable follower configurations before the defender acts. As shown in Appendices I, J, and K, a carefully constructed adversarial sequence can repeatedly mislead the defender, potentially leading to linear regret. This motivates combining the exact oracle with randomized perturbations through the FPL framework.

No-Regret Learning Framework with Perturbations. To mitigate this vulnerability, we adopt an online decision-making approach based on the FPL framework (Kalai & Vempala, 2005). The main idea is to introduce stochastic perturbations to the empirical count matrix $\hat{\mathbf{M}}(t)$, thereby smoothing the leader's decision boundary and preventing adversary from predicting defender's actions. Concretely, at each round t, we perturb the matrix as $\tilde{\mathbf{M}}(t) = \frac{t-1}{t}\hat{\mathbf{M}}(t-1) + \boldsymbol{\epsilon}(t)$, where $\boldsymbol{\epsilon}(t) \sim \text{Uniform}[0, 1/(\delta\sqrt{t})]^{K\times F}$. Then we solve the resulting perturbed optimization problem using our proposed MILP formulation F as an oracle. This integration of online perturbation and MILP-based optimization achieves sublinear expected regret against arbitrarily evolving, non-anticipating attacker sequences. The details are summarized in Algorithm 1.

Define the *cumulative regret* as the difference between the cumulative utility of the best fixed strategy in hindsight and that obtained by the algorithm, written as: $R(T) = \max_{\mathbf{w}\in\mathcal{W}} \sum_{t=1}^{T}(\sum_{k=1}^{K}\sum_{l=1}^{F}\big(\mathbf{M}(t)\odot\mathbf{F}(\mathbf{w})\big)_{kl}) - \sum_{t=1}^{T}(\sum_{k=1}^{K}\sum_{l=1}^{F}\big(\mathbf{M}(t)\odot\mathbf{F}(\mathbf{w}(t))\big)_{kl})$

Theorem 1 (Regret bound for Algorithm 1)**.** *Let $\{\mathbf{w}(t)\}_{t=1}^{T}$ be the sequence of strategies generated by Algorithm 1 with parameter $\delta > 0$ and perturbations $\boldsymbol{\epsilon}(t) \sim \mathcal{U}[0, 1/(\delta\sqrt{t})]^{K\times F}$. For any non-anticipating sequence of follower count matrices $\{\mathbf{M}(t)\}_{t=1}^{T}$ with $n(t) \le C$, the expected cumulative regret satisfies $\mathbb{E}[R(T)] \le \left(4\delta FC^2 + \frac{KF(F+1)}{\delta}\right)\sqrt{T}$. In particular, choosing $\delta = \frac{\sqrt{K(F+1)}}{2C}$ yields $\mathbb{E}[R(T)] \le 4CF\sqrt{K(F+1)T}$.*

Algorithm 1 FPL with an MILP oracle for time-varying-intensity RSSGs

Input: T, F, K, δ and utility parameters. Initialize $\hat{\mathbf{M}}(0) = \mathbf{0}$.

1. For $t = 1, \ldots, T$:
   (a) A non-anticipating environment fixes the hidden $\mathbf{M}(t)$ using the past. For $t > 1$, set $\hat{\mathbf{M}}(t-1) = (t-1)^{-1}\sum_{s<t}\mathbf{M}(s)$.
   (b) Sample $\boldsymbol{\epsilon}(t) \sim \mathcal{U}[0, 1/(\delta\sqrt{t})]^{K\times F}$ and set $\tilde{\mathbf{M}}(t) = \frac{t-1}{t}\hat{\mathbf{M}}(t-1) + \boldsymbol{\epsilon}(t)$.
   (c) Solve the MILP (F.1) using $\tilde{\mathbf{M}}(t)$ and play the maximizing strategy $\mathbf{w}(t)$.
   (d) Followers observe $\mathbf{w}(t)$ and play their prescribed best responses. Observe the full configuration $\mathbf{M}(t)$ after the round.

Moreover, oracle F is not limited to unit-sum allocation ($\sum w_j = 1$) but can naturally generalize to scenarios where the leader can distribute defense budget greater than one, simply by modifying the simplex constraint accordingly. This flexibility allows the proposed framework to handle a wide range of security allocation problems with multiple attacker types and multi-target attacks efficiently.

5 LEARNING UNDER PARTIAL INFORMATION

In our partial information back setting, we consider a setting with at most C attackers acting simultaneously in each round. The follower sequence is arbitrary but fixed before play begins. Under bandit feedback, the defender observes only the aggregate number of attacks on each target, while the number, types, and attack intensities of followers remain unobserved. The methods of Balcan et al. (2015); Harris et al. (2024); Balcan et al. (2025) study single-follower interactions, whereas Personnat et al. (2025) study multiple followers under stochastic type distributions. To our knowledge, prior work has not explicitly addressed bandit learning with simultaneous multiple followers, arbitrarily varying attack intensities, and aggregate feedback.

We therefore aim to develop a no-regret algorithm for the above adversarial multi-follower setting using only aggregate feedback. Specifically, our goal is to achieve $\mathcal{O}(\sqrt{T})$ expected regret for fixed game parameters, without the additional horizon dependence in the guarantees.

5.1 BANDIT MODEL AND FINITE CANDIDATE STRATEGIES

Let $h^k_{j,l}(\mathbf{w})$ denote the best-response indicator defined in Section 3. After committing to $\mathbf{w}(t)$, the defender observes the aggregate attack counts $\nu_j(t) = \sum_{k=1}^{K}\sum_{l=1}^{F} m_{kl}(t)h^k_{j,l}(\mathbf{w}(t))$, where $j = 1, \ldots, N$. The round-t payoff can be written as $r_t(\mathbf{w}(t)) = \sum_{k=1}^{K}\sum_{l=1}^{F} m_{kl}(t)f^k_l(\mathbf{w}(t)) = \sum_{j=1}^{N}\nu_j(t)\left[\lambda^r_j + (\rho^r_j - \lambda^r_j)w_j(t)\right]$. In each round, at most C followers attack up to CF targets. Since the defender's utility from each attack lies in $[-1, 1]$, the total payoff satisfies $|r_t(\mathbf{w})| \le CF$. We use the regret definition in Section 4, with expectations taken over the learner's randomization.

Structured Decision Set. For a fixed follower response profile, the associated defender payoff is linear in $\mathbf{w}$. This motivates constructing a finite candidate set from the vertices of best-response polytopes. For a profile $\sigma = (\sigma^k_l)_{k,l}$ with $\sigma^k_l \in \{0, 1\}^N$ and $\sum_j \sigma^k_{j,l} = l$, define

$$P(\sigma) = \left\{\mathbf{w} \in \mathcal{W} : u^k_a(\mathcal{E}_i, \mathbf{w}) \ge u^k_a(\mathcal{E}_j, \mathbf{w}) \text{ whenever } \sigma^k_{i,l} = 1,\ \sigma^k_{j,l} = 0, \quad \text{for all } k, l\right\}. \tag{5.1}$$

These polytopes cover $\mathcal{W}$. Let

$$V = \bigcup_{\sigma:\, P(\sigma)\neq\varnothing} \operatorname{vert}(P(\sigma)). \tag{5.2}$$

The following lemma shows that, under the prescribed optimistic tie-breaking rule, restricting the defender to V preserves the best fixed payoff exactly. To see why V suffices, take any $\mathbf{w} \in \mathcal{W}$ and let σ be its follower response profile. Keeping σ fixed, the cumulative defender payoff is linear and attains a maximum at some vertex of $P(\sigma)$. At that vertex, σ remains a best response, although other profiles may also become optimal. Optimistic tie-breaking rule chooses σ' with highest defender's expected utility as the attacker's best response. Thus, the best fixed strategy can be chosen from V without any loss in cumulative payoff, establishing the exact reduction below.

Lemma 1 (Exact strategy reduction)**.** *For follower sequence under partial information feedback,*

$$\max_{\mathbf{w}\in\mathcal{W}} \sum_{t=1}^{T} r_t(\mathbf{w}) = \max_{\mathbf{w}\in V} \sum_{t=1}^{T} r_t(\mathbf{w}). \tag{5.3}$$

Moreover, V is independent of T and satisfies $|V| \le \binom{N+K\binom{N}{2}}{N-1}$.

The proof is given in Appendix D.1.

The lemma reduces continuous set $\mathcal{W}$ to finite set V, but V may still be large. For each $\mathbf{w} \in V$, the matrix $F(\mathbf{w})$ records its payoff against every attacker type and attack intensity. The followers present in a round determine how these payoffs are combined, while $F(\mathbf{w})$ remains fixed. The

payoff matrices of all candidates span a space of dimension at most KF. We therefore choose basis strategies from V whose payoff matrices span this space. Each candidate's payoff matrix is a known linear combination of the basis payoff matrices. Since the round payoff is linear in these payoff matrices, the same coefficients express its payoff in terms of the basis strategies' payoffs. By incorporating randomized exploration over the basis strategies, we construct unbiased estimates of their payoffs from the single aggregate payoff observed in each round. Combining these estimates with the known coefficients then yields unbiased payoff estimates for all candidates.

5.2 Payoff coordinates and the EXP2 algorithm

Define $\mathbf{a}(\mathbf{w}) = \mathrm{vec}(F(\mathbf{w})) \in \mathbb{R}^{KF}$, where $d = \dim \mathrm{span}\{\mathbf{a}(\mathbf{w}) : \mathbf{w} \in V\} \le KF$. Apply the same vectorization for $M(t)$. By the barycentric-spanner existence result of Awerbuch & Kleinberg (2008), the finite payoff-feature set admits a basis $B = \{\mathbf{a}(\mathbf{w}_1), \dots, \mathbf{a}(\mathbf{w}_d)\}$ such that every feature has coordinates in $[-1, 1]^d$. Put

$$B = [\,\mathbf{a}(\mathbf{w}_1)\ \cdots\ \mathbf{a}(\mathbf{w}_d)\,], \qquad \mathbf{c}(\mathbf{w}) = (B^\top B)^{-1} B^\top \mathbf{a}(\mathbf{w}). \tag{5.4}$$

Since the columns of B form a basis, these coordinates are unique. Spanner property therefore gives,

$$\mathbf{a}(\mathbf{w}) = B\mathbf{c}(\mathbf{w}), \qquad \|\mathbf{c}(\mathbf{w})\|_\infty \le 1, \qquad \mathbf{c}(\mathbf{w}_i) = \mathbf{e}_i. \tag{5.5}$$

Each basis element is the payoff feature of a complete strategy $\mathbf{w}_i$, so it can be sampled by playing that strategy. Define the normalized loss and the unobserved vector of basis losses by

$$\ell_t(\mathbf{w}) = -\frac{r_t(\mathbf{w})}{CF}, \qquad \mathbf{g}_t = (\ell_t(\mathbf{w}_1), \dots, \ell_t(\mathbf{w}_d))^\top = -\frac{B^\top \mathrm{vec}(M(t))}{CF}. \tag{5.6}$$

The loss vector of strategy $\mathbf{w}$ can be written as: $\ell_t(\mathbf{w}) = \mathbf{c}(\mathbf{w})^\top \mathbf{g}_t$, where $|\ell_t(\mathbf{w})| \le 1$. This equation provides the second reduction. Although best responses make the payoff map discontinuous in $\mathbf{w}$, the losses of all strategies in V depend on the unobserved follower sequence through only $d \le KF$ basis losses. The payoff from any played $\mathbf{w} \in V$ is one linear observation of these losses, with known coefficients $\mathbf{c}(\mathbf{w})$. Let μ be uniform over $\mathbf{w}_1, \dots, \mathbf{w}_d$, and set $p_t = (1-\gamma)x_t + \gamma\mu$, where x_t is the exponential-weights distribution. Since $\mathbf{c}(\mathbf{w}_i) = \mathbf{e}_i$, the second-moment matrix S_t is written as:

$$S_t = \sum_{\mathbf{w} \in V} p_t(\mathbf{w})\mathbf{c}(\mathbf{w})\mathbf{c}(\mathbf{w})^\top = (1-\gamma) \sum_{\mathbf{w} \in V} x_t(\mathbf{w})\mathbf{c}(\mathbf{w})\mathbf{c}(\mathbf{w})^\top + \frac{\gamma}{d} I_d. \tag{5.7}$$

Thus, S_t is invertible and computable from the known coordinates and p_t. After sampling $\mathbf{w}(t) \sim p_t$ and observing its payoff, we have estimators: $\widehat{\mathbf{g}}_t = S_t^{-1}\mathbf{c}(\mathbf{w}(t))\ell_t(\mathbf{w}(t))$ and $\widehat{\ell}_t(\mathbf{w}) = \mathbf{c}(\mathbf{w})^\top \widehat{\mathbf{g}}_t$. Every played strategy, including one sampled through the exploitation part of p_t, contributes to this estimator. Because the basis strategies can actually be played, mixing with μ gives $S_t \succeq (\gamma/d) I_d$. Thus $\mathbb{E}_t[\widehat{\mathbf{g}}_t] = \mathbf{g}_t$, and $\ell_t(\mathbf{w}) = \mathbf{c}(\mathbf{w})^\top \mathbf{g}_t$ gives unbiased loss estimates for every strategy in V. The estimator uses the observation from every round without identifying individual followers.

Algorithm 2 instantiates the EXP2 algorithm (Bubeck et al., 2012) in payoff-basis coordinates. All estimates are computable from aggregate feedback and defender's utility matrix.

5.3 Regret guarantee

Theorem 2 (Square-root regret under aggregate feedback)**.** *Consider the model of Section 5.1, with V defined by* (5.2) *and a basis satisfying* (5.5)*. Suppose $d \ge 1$ and $|V| \ge 2$. For any $\eta > 0$, $0 < \gamma < 1$ with $\eta d^2 \le \gamma$, Algorithm 2 satisfies $\mathbb{E}[R(T)] \le CF\left(\frac{\log|V|}{\eta} + \frac{\eta d T}{1-\gamma} + 2\gamma T\right)$. In particular, set*

$$\eta = \min\left\{\frac{1}{2d^2}, \frac{1}{2d}\sqrt{\frac{\log|V|}{T}}\right\}, \qquad \gamma = \eta d^2. \tag{5.8}$$

Then, for every $T \ge 1$,

$$\mathbb{E}[R(T)] \le 4CFd\sqrt{T \log|V|}. \tag{5.9}$$

The proof is given in Appendix D: Appendix D.2 proves conditional unbiasedness and the relevant second-moment bounds, and Appendix D.3 proves the exponential-weights inequality.

Algorithm 2 Barycentric-spanner EXP2 with aggregate attack feedback

Input: Utility parameters; N, K, F, C, T; $\eta > 0$ and $\gamma \in (0, 1)$.

1. Construct V by (5.2), compute its payoff features, and determine d. If $d = 0$ or $|V| = 1$, play any $\mathbf{w} \in V$ in every round and stop.
2. Compute B and $\mathbf{c}(\mathbf{w})$ satisfying (5.5). Let μ be uniform over the basis strategies, and initialize $x_1(\mathbf{w}) = 1/|V|$ for all $\mathbf{w} \in V$.
3. **For** $t = 1, \ldots, T$:
   (a) Set $p_t = (1 - \gamma)x_t + \gamma\mu$ and compute S_t by (5.7).
   (b) Sample $\mathbf{w}(t) \sim p_t$ and commit to it. Followers then best respond.
   (c) Observe $\nu(t)$, compute $\ell_t(\mathbf{w}(t)) = -r_t(\mathbf{w}(t))/(CF)$, and estimate all candidate losses.
   (d) Update $x_{t+1}(\mathbf{w}) \propto x_t(\mathbf{w}) \exp(-\eta\widehat{\ell}_t(\mathbf{w}))$ for all $\mathbf{w} \in V$, normalized to sum to one.

The aggregate payoff observed in each round is used to estimate the losses of all candidate strategies. Crucially, the weighted conditional second moment satisfies $\mathbb{E}_t[\sum_{\mathbf{w}\in V} x_t(\mathbf{w})\widehat{\ell}_t(\mathbf{w})^2] \le d/(1-\gamma)$. With $\gamma = \eta d^2 \le 1/2$, the second-moment and exploration terms satisfy $\eta dT/(1-\gamma)+2\gamma T \le 4d^2\eta T$. Substituting this into the general regret bound yields $\mathbb{E}[R(T)] \le CF\left(\frac{\log|V|}{\eta} + 4d^2\eta T\right)$. The parameter choice in (5.8) yields the square-root bound. In particular, exploration is paid for through $2\gamma T$, while the learner updates after every round rather than waiting for a block of basis observations.

To sharpen the dependence on the effective dimension d, we further optimize the exploration distribution while retaining the same loss estimator and exponential weights update. This yields tighter control of the loss estimates and permits $\gamma = \eta d$ instead of $\gamma = \eta d^2$. With an appropriate learning rate, the resulting regret bound becomes $\mathbb{E}[R(T)] \le 4CF\sqrt{dT\log|V|}$, improving the bound in Theorem 2 by a factor of $\sqrt{d}$. The optimized exploration distribution is computed once from the known payoff features; the construction and proof are provided in Appendix D.5 (Proposition 3).

We next establish a lower bound on the dependence of the regret on the horizon. The key idea is to construct a simple RSSG instance in which the payoff difference between two fixed defender strategies can be represented by a symmetric random walk. Specifically, the two possible attack intensities induce independent $\{-1, +1\}$ increments, so that the cumulative advantage of the hindsight-optimal fixed strategy is governed by the fluctuation of the resulting random walk. Since this fluctuation is of order $\sqrt{T}$, the construction yields the following lower bound.

Proposition 2 (Lower bound in the horizon)**.** *For every $C \ge 1$ and $T \ge 1$, there exists an RSSG instance with $N = 2$, $K = 1$, and $F = 2$ such that, for every possibly randomized defender algorithm, there exists an oblivious sequence $\{M(t)\}_{t=1}^{T}$ satisfying $\sum_{k=1}^{K}\sum_{\ell=1}^{F} m_{k\ell}(t) = C$, for which*

$$\mathbb{E}[R(T)] \ge \frac{C}{4\sqrt{3}}\sqrt{T}. \tag{5.10}$$

The expectation is over the learner's internal randomization. Moreover, the lower bound continues to hold even if the learner receives full-information feedback after every round.

Proposition 2 and Theorem 2 together show that the $\sqrt{T}$ dependence on the horizon is order-optimal.

Experimental evaluation. We evaluate Algorithm 1 under adversarial and stochastic follower sequences, and Algorithm 2 under aggregate feedback with time-varying attacker sequence. The experimental settings and results are reported in Appendix A.

6 Conclusion

We study no-regret learning in repeated Stackelberg security games with multiple followers and time-varying attack intensities. Under full-information feedback, an exact MILP oracle supports FPL without explicit strategy enumeration. Under bandit feedback, barycentric-spanner EXP2 constructs unbiased estimates of all candidate payoffs from a single payoff per round. In our game model, both algorithms achieve $O(\sqrt{T})$ expected regret against adversarial follower sequences. These bounds contain no additional logarithmic factors in T and match the $\Omega(\sqrt{T})$ lower bound. Numerical experiments illustrate the performance of both algorithms.

AI USE STATEMENT

Generative AI tools were used for language editing, checking mathematical proofs for potential errors and gaps, and improving the presentation of mathematical expressions. They also assisted with developing code for the runtime of LP enumeration baseline. The main research ideas were developed by the authors, who retain full responsibility for the mathematical arguments, experimental results, and final manuscript.

REPRODUCIBILITY STATEMENT

The game and feedback models, together with the round protocols, are specified in Sections 2, 4, and 5.1. Algorithms 1 and 2 describe the learning procedures, and the complete MILP formulations appear in Appendix F. Proofs of the regret guarantees, including the analysis of the optimized exploration distribution, are provided in Appendices C and D. The simulation settings and parameter choices are described in Appendices H–L. The runtime comparison, including the computational environment and evaluation protocol, is documented in Appendix G.

REFERENCES

Manos Antonakakis, Tim April, Michael Bailey, Matt Bernhard, Elie Bursztein, Jaime Cochran, Zakir Durumeric, J Alex Halderman, Luca Invernizzi, Michalis Kallitsis, et al. Understanding the mirai botnet. In *26th USENIX security symposium (USENIX Security 17)*, pp. 1093–1110, 2017.

Baruch Awerbuch and Robert Kleinberg. Online linear optimization and adaptive routing. *Journal of Computer and System Sciences*, 74(1):97–114, 2008.

Francesco Bacchiocchi, Matteo Bollini, Matteo Castiglioni, Alberto Marchesi, and Nicola Gatti. The sample complexity of stackelberg games. *arXiv preprint arXiv:2405.06977*, 2024.

Rufan Bai, Haoxing Lin, Xinyu Yang, Xiaowei Wu, Minming Li, and Weijia Jia. Stackelberg security games with contagious attacks on a network: Reallocation to the rescue. *Journal of Artificial Intelligence Research*, 77:487–515, 2023.

Maria-Florina Balcan, Avrim Blum, Nika Haghtalab, and Ariel D. Procaccia. Commitment without regrets: Online learning in Stackelberg security games. In *16th ACM Conf. on Economics and Computation*, pp. 61–78, 2015.

Maria-Florina Balcan, Martino Bernasconi, Matteo Castiglioni, Andrea Celli, Keegan Harris, and Zhiwei Steven Wu. Nearly-optimal bandit learning in stackelberg games with side information. *arXiv preprint arXiv:2502.00204*, 2025.

Matthew Becker, Rachel McRobb, Fred Watson, Egil Droge, Benson Kanyembo, James Murdoch, and Catherine Kakumbi. Evaluating wire-snare poaching trends and the impacts of by-catch on elephants and large carnivores. *Biological conservation*, 158:26–36, 2013.

Martino Bernasconi, Matteo Castiglioni, Andrea Celli, Alberto Marchesi, Francesco Trovò, and Nicola Gatti. Optimal rates and efficient algorithms for online bayesian persuasion. In *Intl. Conf. on Machine Learning*, pp. 2164–2183, 2023.

Avrim Blum, Nika Haghtalab, and Ariel D Procaccia. Learning optimal commitment to overcome insecurity. In *Proc. Advances in Neural Information Processing Systems*, volume 27, 2014.

Michael Brückner and Tobias Scheffer. Stackelberg games for adversarial prediction problems. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 547–555, 2011.

Sébastien Bubeck, Nicolò Cesa-Bianchi, and Sham M. Kakade. Towards minimax policies for online linear optimization with bandit feedback. *arXiv preprint arXiv:1202.3079*, 2012. URL https://arxiv.org/abs/1202.3079.

Matteo Castiglioni, Andrea Celli, Alberto Marchesi, and Nicola Gatti. Online bayesian persuasion. *Advances in neural information processing systems*, 33:16188–16198, 2020.

Guanda Chen, Yiding Ji, and Shuo Han. Lightweight no-regret online learning in repeated stackelberg security games. *IEEE Control Systems Letters*, 2026.

Xu Chen, Liang Xiao, Wei Feng, Ning Ge, and Xianbin Wang. Ddos defense for iot: A stackelberg game model-enabled collaborative framework. *IEEE Internet of Things Journal*, 9(12):9659–9674, 2021.

Zhaoyang Cheng, Guanpu Chen, and Yiguang Hong. Single-leader-multiple-followers stackelberg security game with hypergame framework. *IEEE Transactions on Information Forensics and Security*, 17:954–969, 2022.

Vincent Conitzer and Tuomas Sandholm. Computing the optimal strategy to commit to. In *Proceedings of the 7th ACM conference on Electronic commerce*, pp. 82–90, 2006.

Kate Donahue, Nicole Immorlica, Meena Jagadeesan, Brendan Lucier, and Aleksandrs Slivkins. Impact of decentralized learning on player utilities in stackelberg games. *arXiv preprint arXiv:2403.00188*, 2024.

Jiarui Gan, Edith Elkind, Sarit Kraus, and Michael Wooldridge. Defense coordination in security games: Equilibrium analysis and mechanism design. *Artificial Intelligence*, 313:103791, 2022.

Gurobi Optimization, LLC. Gurobi Optimizer Reference Manual, 2024.

Keegan Harris, Steven Z Wu, and Maria-Florina F Balcan. Regret minimization in stackelberg games with side information. *Advances in Neural Information Processing Systems*, 37:12944–12976, 2024.

Grant Ho, Mayank Dhiman, Devdatta Akhawe, Vern Paxson, Stefan Savage, Geoffrey M Voelker, and David Wagner. Hopper: Modeling and detecting lateral movement. In *30th USENIX Security Symposium (USENIX Security 21)*, pp. 3093–3110, 2021.

Adam Kalai and Santosh Vempala. Efficient algorithms for online decision problems. *Journal of Computer and System Sciences*, 71(3):291–307, 2005.

Joshua Letchford, Vincent Conitzer, and Kamesh Munagala. Learning and approximating the optimal strategy to commit to. In *International symposium on algorithmic game theory*, pp. 250–262. Springer, 2009.

Tao Li and Suresh P Sethi. A review of dynamic stackelberg game models. *Discrete & Continuous Dynamical Systems-Series B*, 22(1), 2017.

Douglas C MacMillan and Quoc Anh Nguyen. Factors influencing the illegal harvest of wildlife by trapping and snaring among the katu ethnic group in vietnam. *Oryx*, 48(2):304–312, 2014.

Praveen Paruchuri, Jonathan P. Pearce, Janusz Marecki, Milind Tambe, Fernando Ordonez, and Sarit Kraus. Playing games for security: An efficient exact algorithm for solving bayesian Stackelberg games. In *7th Intl. Joint Conference on Autonomous Agents and Multiagent Systems*, pp. 895–902, 2008.

Gerson Personnat, Tao Lin, Safwan Hossain, and David C Parkes. Learning to play multi-follower bayesian stackelberg games. *arXiv preprint arXiv:2510.01387*, 2025.

James Pita, Manish Jain, Janusz Marecki, Fernando Ordóñez, Christopher Portway, Milind Tambe, Craig Western, Praveen Paruchuri, and Sarit Kraus. Deployed ARMOR protection: the application of a game theoretic model for security at the Los Angeles International Airport. In *7th International Joint Conference on Autonomous Agents and Multiagent Systems*, pp. 125–132, 2008.

Yundi Qian, Chao Zhang, Bhaskar Krishnamachari, and Milind Tambe. Restless poachers: Handling exploration-exploitation tradeoffs in security domains. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*, pp. 123–131, 2016.

Eric Shieh, Bo An, Rong Yang, Milind Tambe, Craig Baldwin, Joseph DiRenzo, Ben Maule, and Garrett Meyer. Protect: A deployed game theoretic system to protect the ports of the united states. In *Proc. 11th International Conference on Autonomous Agents and Multiagent Systems*, pp. 13–20, 2012.

Omar Abdel Wahab, Jamal Bentahar, Hadi Otrok, and Azzam Mourad. Resource-aware detection and defense system against multi-type attacks in the cloud: Repeated bayesian stackelberg game. *IEEE Transactions on Dependable and Secure Computing*, 18(2):605–622, 2019.

Shiyang Xu, Yongqiang Guan, and Yu Shen. A stackelberg game for optimal control energy of multi-agent networks under cyber-attacks. *International Journal of Control*, 99(1):164–175, 2026.

Rong Yang, Benjamin J. Ford, Milind Tambe, and Andrew Lemieux. Adaptive resource allocation for wildlife protection against illegal poachers. In *13rd Intl. Joint Conf. on Auto. Agents and Multiagent Systems*, pp. 453–460, 2014.

Peng Yu, Yanpeng Hu, Yinghui Wang, Ruizhe Jia, and Jin Guo. Optimal consensus control strategy for multi-agent systems under cyber attacks via a stackelberg game approach. *IEEE Transactions on Automation Science and Engineering*, 2025.

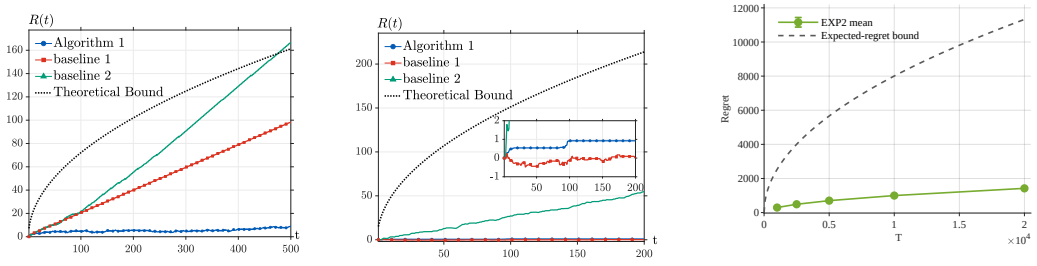
Mengxiao Zhang, Peng Zhao, Haipeng Luo, and Zhi-Hua Zhou. No-regret learning in time-varying zero-sum games. In *International Conference on Machine Learning*, pp. 26772–26808, 2022.

A SIMULATIONS

The following experiments assess the Algorithm 1 and Algorithm 2. The guarantee for the per-round EXP2 algorithm is established theoretically in Theorem 2.

Figure 1(a) uses $C = 2$, $K = 2$, and $F = 2$ with the adversarial attacker sequence described in Appendix K. The oracle-based Follow-the-Leader baseline can be exploited by the sequence. The second baseline, Chen et al. (2026), does not account for varying attack intensity. Figure 1(b) considers $N = 5$, $K = 5$, $F = 3$, and $C = 2$ with the stochastic attacker distributions described in Appendix H. The additional full-information simulations and their settings remain in Appendices I–L.

Figure 1(c) considers $N = K = F = C = 2$ with the alternating attacker distributions described in Appendix M. The type and intensity distributions change every 250 rounds while retaining the same unique expected-optimal strategy. Barycentric-spanner EXP2 uses only the aggregate payoff of the selected strategy. The green markers show mean terminal regret from separate runs at each total horizon, with parameters selected for that horizon and fixed throughout each run. The dashed curve gives the corresponding theoretical $\mathcal{O}(\sqrt{T})$ upper bound. These results illustrate learning from aggregate feedback under changing attacker composition.



(a) Multiple follower with adversarial Attacker sequence

(b) Multiple followers with general distribution of attackers

(c) barycentric-spanner EXP2 under partial information feedback

Figure 1: Retained cumulative-regret plots: (a) adversarial and (b) stochastic full-information settings; (c) barycentric-spanner EXP2.

B COMPUTATIONAL COMPLEXITY OF THE MILP ORACLE

Proof of Proposition 1. For a fixed attacker type, the binary vectors $\mathbf{h}_1, \ldots, \mathbf{h}_F$ are nested and have cardinalities $1, \ldots, F$. Each transition adds one previously unselected target. Hence every binary assignment is specified by an ordered list of F distinct targets, giving $N!/(N-F)!$ possibilities. Across the K types there are at most $\big(N!/(N-F)!\big)^K$ combinations. For each combination, fix the binary variables. The remaining optimization problem is a linear program in $\mathbf{w}$, the $\mathbf{z}_l^k$, and the thresholds τ_l^k, with $O(KFN)$ variables and constraints. Solve each feasible linear program and choose the solution with the largest objective value. Every feasible binary assignment has been considered, so the resulting solution is globally optimal. This proves the enumeration bound. A general linear program is not assumed to admit a linear-time solver; its actual solution cost must be included in the runtime bound. □

C PROOF OF THE FULL-INFORMATION REGRET BOUND

We consider time-varying follower numbers with follower number bounded by C and arbitrarily attacker types from $[K]$ and F attack intensities of each follower. Let the state at round t be $\mathbf{M}(t)$, where each element indicates the number $m_{kl}(t)$ of active attacker α_k with specific attack intensity l.

The objective over leader decision $\mathbf{w}$ is summarized as

$$\sum_{k=1}^{K}\sum_{l=1}^{F}\big(\hat{\mathbf{M}}(t)\odot\mathbf{F}(\mathbf{w})\big)_{kl},$$

For each active attacker with type α_k attack intensity l, the utility term is written as $\big(\mathbf{F}(\mathbf{w})\big)_{kl} = f_l^k(\mathbf{w})$ defined in section 4. Function $f_l^k(\mathbf{w})$ is non-convex and discontinuous. For $\mathbf{b}_k(\mathbf{w})$ with $|\mathbf{b}_k(\mathbf{w})| = l$, discontinuities appear when the attacker's best response changes. Hence $f_l^k(\mathbf{w})$ is divided into multiple regions and is piecewise linear (conditioned on $\mathbf{b}_k(\mathbf{w})$).

The defender strategy at time t is modeled by

$$\mathbf{w}(t) = \arg\max_{\mathbf{w}} \sum_{k=1}^{K}\sum_{l=1}^{F}\big((\frac{t-1}{t}\hat{\mathbf{M}}(t-1) + \boldsymbol{\epsilon}(t)) \odot \mathbf{F}(\mathbf{w})\big)_{kl},,$$

where each attacker type with specific attack intensity corresponds to a distinct piecewise-linear utility function, and $\hat{\mathbf{M}}(t-1)$ is the attack-intensity frequency (state-frequency) vector. We have $(\mathbf{F}(\mathbf{w}))_{kl} = f_l^k(\mathbf{w})$, and write instantaneous regret as

$$r(t) = \sum_{k=1}^{K}\sum_{l=1}^{F}[\big(\mathbf{M}(t) \odot \mathbf{F}(\mathbf{w}^*(T))\big)_{kl} - \big(\mathbf{M}(t) \odot \mathbf{F}(\mathbf{w}(t))\big)_{kl}].$$

Therefore,

$$\mathbb{E}[R_T] = \mathbb{E}\left[\sum_{t=1}^{T}\sum_{k=1}^{K}\sum_{l=1}^{F}[\big(\mathbf{M}(t) \odot \mathbf{F}(\mathbf{w}^*(T))\big)_{kl} - \big(\mathbf{M}(t) \odot \mathbf{F}(\mathbf{w}(t))\big)_{kl}]\right],$$

where equation (F.1) is the oracle in FTPL.

Linear Transformation The online optimization problem at each round t is written as:

$$\max_{\mathbf{F}(\mathbf{w})} \sum_{\tau=1}^{t}\sum_{k=1}^{K}\sum_{l=1}^{F}\big(\mathbf{M}(\tau) \odot \mathbf{F}(\mathbf{w})\big)_{kl}.$$

We reformulate this problem by introducing a decision variable matrix $\mathbf{Y} = [y_l^k]_{k=1,\dots,K}^{l=1,\dots,F}$, where each component $y_l^k = f_l^k(\mathbf{w}) = \sum_{\mathcal{E}_j \in \mathbf{b}_k(\mathbf{w})} \lambda_j^r(1-w_j) + \rho_j^r w_j$ represents the expected payoff of the defender against the attacker α_k. Stated as:

$$\begin{aligned} \max_{\mathbf{Y}} \quad & \langle \mathbf{Y}, \hat{\mathbf{M}}(t)\rangle_F, \\ \text{s.t.} \quad & \mathbf{Y} \in \mathcal{Y}, \end{aligned} \tag{C.1}$$

where $\langle \mathbf{Y}, \hat{\mathbf{M}}(t)\rangle_F$ represents Frobenius inner product and $\mathcal{Y} = \mathbf{F}(\mathcal{W})$ is the image of the original feasible region under the mapping $\mathbf{F}$ defined in section 4. Although the original utility vector function of the defender $\mathbf{F}(\mathbf{w})$ is non-linear due to the best-response structure, our reformulated objective function becomes linear in terms of the new variable $\mathbf{Y}$ and is subject to a non-convex feasible set.

Additionally, the above linear reformulation allows us to leverage the FPL framework to ensure a sublinear regret bound. Notably, FPL achieves sublinear regret for linear objectives over any decision set including non-convex ones, contingent on the existence of efficient optimization oracles. The MILP formulation (3.3) just serves as such an oracle required by FPL (Kalai & Vempala, 2005). We now recall a key result from (Kalai & Vempala, 2005) and it is instrumental to derive the regret bound of our approach.

Let $\mathbb{E}[\mathcal{R}_T(\delta)]$ denote the expected regret of $\mathbf{Hanna}(\delta)$ up to time T. Follow Hannan lead and use gradually increasing perturbations:

Lemma 2. *(Kalai & Vempala, 2005)* $\mathbf{Hanna}(\delta)$*: on each period t,*

1. *Choose* $\boldsymbol{\epsilon}(t)$ *uniformly at random from the cube* $[0, 1/\delta\sqrt{t}]^{KF}$*;*
2. *Use strategy* $\arg\max_{\mathbf{Y}} \langle \mathbf{Y}, \tilde{\mathbf{M}}(t)\rangle$.

For any state matrix sequence $\mathbf{M}(1), \mathbf{M}(2), \dots$*, after any number of periods* $T > 0$*,*

$$\mathbb{E}[\mathcal{R}_T(\delta)] \leq 2\delta RG\sqrt{T} + \frac{D\sqrt{T}}{\delta}, \tag{C.2}$$

where the constants are defined as:

$$D := \sup_{\mathbf{w},\mathbf{w}' \in \mathcal{W}} \|\mathbf{F}(\mathbf{w}) - \mathbf{F}(\mathbf{w}')\|_1,$$
$$R := \sup_{\substack{\mathbf{w},\mathbf{w}' \in \mathcal{W},\ \mathbf{M} \geq 0 \\ \sum_{k,l} m_{kl} \leq C}} |\langle \mathbf{M}, \mathbf{F}(w) - \mathbf{F}(w') \rangle_{\mathrm{F}}|,$$
$$G := \max_t \|\mathbf{M}(t)\|_1.$$

In our setting, the constants can be bounded as follows. Since each component $f_l^k(\mathbf{w})$ lies in $[-l, l]$, we have $|f_l^k(\mathbf{w}) - f_l^k(\mathbf{w}')| \leq 2l$. Summing over all $k \in [K]$ and $l \in [F]$, it follows that

$$D = \sup_{\mathbf{w},\mathbf{w}' \in \mathcal{W}} \|\mathbf{F}(\mathbf{w}) - \mathbf{F}(\mathbf{w}')\|_1 \leq \sum_{k=1}^{K} \sum_{l=1}^{F} 2l = KF(F+1).$$

Similarly, $|f_l^k(\mathbf{w}) - f_l^k(\mathbf{w}')| \leq 2F$ and each round contains at most C followers, the one-round payoff range satisfies

$$R := \sup_{\substack{\mathbf{w},\mathbf{w}' \in \mathcal{W},\ \mathbf{M} \geq 0 \\ \sum_{k,l} m_{kl} \leq C}} |\langle \mathbf{M}, \mathbf{F}(w) - \mathbf{F}(w') \rangle_{\mathrm{F}}| \leq 2FC.$$

For G, note that $\mathbf{M}(t)$ is a count matrix where each entry $m_{kl}(t)$ represents the number of followers choosing attacker type k with attack intensity l at time t. Since each follower selects exactly one (k, l) pair, the ℓ_1-norm of $\mathbf{M}(t)$ equals the number of followers at time t, denoted by $n(t)$. As $n(t) \leq C$, we obtain

$$G = \max_t \|\mathbf{M}(t)\|_1 = \max_t n(t) \leq C.$$

Substituting these bounds into (C.2), we obtain

$$\mathbb{E}[\mathcal{R}_T(\delta)] \leq 4\delta FC^2\sqrt{T} + \frac{KF(F+1)\sqrt{T}}{\delta}. \tag{C.3}$$

Choosing

$$\delta = \frac{\sqrt{K(F+1)}}{2C}$$

minimizes the right-hand side of (C.3). Substituting this choice gives

$$\mathbb{E}[\mathcal{R}_T(\delta)] \leq 2CF\sqrt{K(F+1)T} + 2CF\sqrt{K(F+1)T} = 4CF\sqrt{K(F+1)T}.$$

Thus, for fixed game parameters, the expected regret is $\mathcal{O}(\sqrt{T})$. Therefore, the regret is $\mathcal{O}(\sqrt{T})$, with constants depending on D, R and G.

Finally, the original bi-level optimization problem can be reduced to a standard single-level MILP. This reformulation enables efficient computation of the optimal defense strategy using off-the-shelf solvers such as Gurobi or CPLEX, making real-time implementation feasible.

D Proofs for the partial-information algorithm

D.1 Exact finite-strategy reduction

Proof of Lemma 1. Fix $\mathbf{w} \in W$ and its realized response profile σ. With σ held fixed, the cumulative defender payoff is linear on $P(\sigma)$ and is therefore maximized at a vertex $\mathbf{v} \in V$. At $\mathbf{v}$, σ remains a best-response profile, and optimistic tie-breaking can only improve the defender's payoff. Hence $\sum_t r_t(\mathbf{v}) \geq \sum_t r_t(\mathbf{w})$, proving (5.3).

For $|V| \leq \binom{N+K\binom{N}{2}}{N-1}$, the region boundaries come from the N equations $w_j = 0$ and at most $K\binom{N}{2}$ pairwise attacker-utility equalities. In the $(N-1)$-dimensional affine space $\sum_j w_j = 1$, each vertex is specified by $N-1$ independent active equalities. Counting their possible choices gives the stated bound. □

D.2 UNBIASEDNESS AND SECOND MOMENTS

Let $\mathcal{F}_{t-1}$ denote the learner's history before its round-t draw, and write $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot \mid \mathcal{F}_{t-1}]$. For the given oblivious sequence, $\mathbf{g}_t$ and $\ell_t(\mathbf{w})$ are fixed, while x_t, p_t, S_t are determined by the history.

Lemma 3 (Per-round payoff estimation)**.** *For every round of Algorithm 2,*

$$S_t \succeq (\gamma/d) I_d, \tag{D.1}$$

$$\mathbb{E}_t[\widehat{\mathbf{g}}_t] = \mathbf{g}_t, \qquad \mathbb{E}_t[\widehat{\ell}_t(\mathbf{w})] = \ell_t(\mathbf{w}) \quad (\mathbf{w} \in V), \tag{D.2}$$

$$|\widehat{\ell}_t(\mathbf{w})| \le d^2/\gamma \quad (\mathbf{w} \in V), \tag{D.3}$$

$$\mathbb{E}_t\left[\sum_{\mathbf{w}\in V} p_t(\mathbf{w})\widehat{\ell}_t(\mathbf{w})^2\right] \le d, \tag{D.4}$$

$$\mathbb{E}_t\left[\sum_{\mathbf{w}\in V} x_t(\mathbf{w})\widehat{\ell}_t(\mathbf{w})^2\right] \le \frac{d}{1-\gamma}. \tag{D.5}$$

Proof. The identity in (5.7) proves (D.1) because its first summand is positive semidefinite. Using the exact loss representation and summing over the actual sampling distribution gives

$$\begin{aligned}\mathbb{E}_t[\widehat{\mathbf{g}}_t] &= S_t^{-1} \sum_{\mathbf{w}\in V} p_t(\mathbf{w})\mathbf{c}(\mathbf{w})\ell_t(\mathbf{w}) \\ &= S_t^{-1} \sum_{\mathbf{w}\in V} p_t(\mathbf{w})\mathbf{c}(\mathbf{w})\mathbf{c}(\mathbf{w})^\top \mathbf{g}_t = S_t^{-1} S_t \mathbf{g}_t = \mathbf{g}_t.\end{aligned} \tag{D.6}$$

Multiplication by the fixed vector $\mathbf{c}(\mathbf{w})^\top$ proves the second equality in (D.2).

By (D.1), $\|S_t^{-1}\|_{\mathrm{op}} \le d/\gamma$. Together with $|\ell_t(\mathbf{w}(t))| \le 1$ and $\|\mathbf{c}(\mathbf{w})\|_2 \le \sqrt{d}$ from (5.5), this gives

$$\begin{aligned}|\widehat{\ell}_t(\mathbf{w})| &= |\mathbf{c}(\mathbf{w})^\top S_t^{-1} \mathbf{c}(\mathbf{w}(t))\ell_t(\mathbf{w}(t))| \\ &\le \|\mathbf{c}(\mathbf{w})\|_2 \|S_t^{-1}\|_{\mathrm{op}} \|\mathbf{c}(\mathbf{w}(t))\|_2 \le \sqrt{d}\,\frac{d}{\gamma}\,\sqrt{d} = \frac{d^2}{\gamma}.\end{aligned} \tag{D.7}$$

For the second-moment estimate, first condition on the chosen strategy and substitute the estimator:

$$\begin{aligned}\sum_{\mathbf{w}\in V} p_t(\mathbf{w})\widehat{\ell}_t(\mathbf{w})^2 &= \widehat{\mathbf{g}}_t^\top \left(\sum_{\mathbf{w}\in V} p_t(\mathbf{w})\mathbf{c}(\mathbf{w})\mathbf{c}(\mathbf{w})^\top\right)\widehat{\mathbf{g}}_t \\ &= \widehat{\mathbf{g}}_t^\top S_t \widehat{\mathbf{g}}_t \\ &= \ell_t(\mathbf{w}(t))^2 \mathbf{c}(\mathbf{w}(t))^\top S_t^{-1} \mathbf{c}(\mathbf{w}(t)).\end{aligned} \tag{D.8}$$

The scalar square is at most one. The identity $z^\top A z = \mathrm{tr}(A z z^\top)$ and the definition of S_t then imply

$$\begin{aligned}\mathbb{E}_t\left[\sum_{\mathbf{w}\in V} p_t(\mathbf{w})\widehat{\ell}_t(\mathbf{w})^2\right] &\le \mathbb{E}_t[\mathbf{c}(\mathbf{w}(t))^\top S_t^{-1} \mathbf{c}(\mathbf{w}(t))] \\ &= \mathrm{tr}\big(S_t^{-1}\mathbb{E}_t[\mathbf{c}(\mathbf{w}(t))\mathbf{c}(\mathbf{w}(t))^\top]\big) \\ &= \mathrm{tr}(S_t^{-1} S_t) = d.\end{aligned} \tag{D.9}$$

Finally, $p_t(\mathbf{w}) \ge (1-\gamma)x_t(\mathbf{w})$ and $\widehat{\ell}_t(\mathbf{w})^2 \ge 0$. Multiplying this pointwise probability inequality by the squared estimates, summing, and taking conditional expectations yields (D.5). □

D.3 THE EXPONENTIAL-WEIGHTS INEQUALITY

Lemma 4 (Exponential weights with signed loss estimates)**.** *Let x_1 be uniform on V, and let*

$$x_{t+1}(\mathbf{w}) = \frac{x_t(\mathbf{w}) e^{-\eta \widehat{\ell}_t(\mathbf{w})}}{\sum_{\mathbf{v}\in V} x_t(\mathbf{v}) e^{-\eta \widehat{\ell}_t(\mathbf{v})}}. \tag{D.10}$$

If $|\eta\widehat{\ell}_t(\mathbf{w})| \le 1$ *for every* $t, \mathbf{w}$*, then for every* $\mathbf{w}^\star \in V$*, pathwise,*

$$\sum_{t=1}^{T}\left(\sum_{\mathbf{w}\in V} x_t(\mathbf{w})\widehat{\ell}_t(\mathbf{w}) - \widehat{\ell}_t(\mathbf{w}^\star)\right) \le \frac{\log|V|}{\eta} + \eta\sum_{t=1}^{T}\sum_{\mathbf{w}\in V} x_t(\mathbf{w})\widehat{\ell}_t(\mathbf{w})^2. \tag{D.11}$$

Proof. We want to compare the average estimated loss under x_t with the estimated loss of a fixed strategy $\mathbf{w}^\star$. The update formula provides a starting point: the numerator is available to contain the estimated loss of $\mathbf{w}^\star$, while the denominator contains the estimated losses of all strategies, weighted by x_t. Taking logarithms separates these two parts and brings $-\eta\widehat{\ell}_t(\mathbf{w}^\star)$ out of the exponential. To obtain the weighted average of the estimated losses from the denominator, we use $e^{-z} \le 1 - z + z^2$ followed by $\log u \le u - 1$.

Since $|\eta\widehat{\ell}_t(\mathbf{w})| \le 1$, applying $e^{-z} \le 1 - z + z^2$ gives

$$\begin{aligned}\sum_{\mathbf{w}\in V} x_t(\mathbf{w})e^{-\eta\widehat{\ell}_t(\mathbf{w})} &\le \sum_{\mathbf{w}\in V} x_t(\mathbf{w})\left(1 - \eta\widehat{\ell}_t(\mathbf{w}) + \eta^2\widehat{\ell}_t(\mathbf{w})^2\right)\\ &= 1 - \eta\sum_{\mathbf{w}\in V} x_t(\mathbf{w})\widehat{\ell}_t(\mathbf{w}) + \eta^2\sum_{\mathbf{w}\in V} x_t(\mathbf{w})\widehat{\ell}_t(\mathbf{w})^2.\end{aligned} \tag{D.12}$$

Then, using $\log u \le u - 1$ for $u > 0$, we obtain

$$\begin{aligned}\log\left(\sum_{\mathbf{w}\in V} x_t(\mathbf{w})e^{-\eta\widehat{\ell}_t(\mathbf{w})}\right) &\le \sum_{\mathbf{w}\in V} x_t(\mathbf{w})e^{-\eta\widehat{\ell}_t(\mathbf{w})} - 1\\ &\le -\eta\sum_{\mathbf{w}\in V} x_t(\mathbf{w})\widehat{\ell}_t(\mathbf{w}) + \eta^2\sum_{\mathbf{w}\in V} x_t(\mathbf{w})\widehat{\ell}_t(\mathbf{w})^2.\end{aligned} \tag{D.13}$$

The update (D.10) also gives

$$\log\frac{x_{t+1}(\mathbf{w}^\star)}{x_t(\mathbf{w}^\star)} = -\eta\widehat{\ell}_t(\mathbf{w}^\star) - \log\left(\sum_{\mathbf{w}\in V} x_t(\mathbf{w})e^{-\eta\widehat{\ell}_t(\mathbf{w})}\right). \tag{D.14}$$

Combining the two displays and rearranging gives

$$\sum_{\mathbf{w}\in V} x_t(\mathbf{w})\widehat{\ell}_t(\mathbf{w}) - \widehat{\ell}_t(\mathbf{w}^\star) \le \frac{1}{\eta}\log\frac{x_{t+1}(\mathbf{w}^\star)}{x_t(\mathbf{w}^\star)} + \eta\sum_{\mathbf{w}\in V} x_t(\mathbf{w})\widehat{\ell}_t(\mathbf{w})^2. \tag{D.15}$$

We now sum this inequality over all rounds, keeping the same comparison strategy $\mathbf{w}^\star$. Summing over t telescopes the logarithms. Since $x_1(\mathbf{w}^\star) = 1/|V|$ and $x_{T+1}(\mathbf{w}^\star) \le 1$,

$$\sum_{t=1}^{T}\log\frac{x_{t+1}(\mathbf{w}^\star)}{x_t(\mathbf{w}^\star)} = \log\frac{x_{T+1}(\mathbf{w}^\star)}{x_1(\mathbf{w}^\star)} \le \log|V|.$$

Substitution proves (D.11).

Although the loss estimates depend on the strategies sampled by the algorithm, the calculations above hold for any realized sequence of estimates satisfying the stated bound. Thus, the inequality holds for every possible sampling path, without taking expectations or requiring independence across rounds. □

D.4 Proof of the regret theorem

Proof of Theorem 2. First consider $d \ge 1$ and $|V| \ge 2$. Because $\eta d^2 \le \gamma$, Lemma 3 ensures $|\eta\widehat{\ell}_t(\mathbf{w})| \le 1$. Thus Lemma 4 applies to every realization of the algorithm. For the given oblivious sequence, choose

$$\mathbf{w}^\star \in \arg\min_{\mathbf{w}\in V}\sum_{t=1}^{T}\ell_t(\mathbf{w}).$$

This comparison strategy is fixed independently of the learner's random draws. By Lemma 1, it is also optimal for the original continuous benchmark.

Step 1: Replace estimated losses by true losses. Conditional on $\mathcal{F}_{t-1}$, the distribution x_t is fixed. Lemma 3 therefore implies

$$\mathbb{E}_t\left[\sum_{\mathbf{w}\in V} x_t(\mathbf{w})\widehat{\ell}_t(\mathbf{w}) - \widehat{\ell}_t(\mathbf{w}^\star)\right] = \sum_{\mathbf{w}\in V} x_t(\mathbf{w})\ell_t(\mathbf{w}) - \ell_t(\mathbf{w}^\star). \tag{D.16}$$

Taking full expectations and summing over t, the tower property gives

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T}\left(\sum_{\mathbf{w}\in V} x_t(\mathbf{w})\widehat{\ell}_t(\mathbf{w}) - \widehat{\ell}_t(\mathbf{w}^\star)\right)\right] \\
&\quad= \sum_{t=1}^{T}\mathbb{E}\left[\mathbb{E}_t\left[\sum_{\mathbf{w}\in V} x_t(\mathbf{w})\widehat{\ell}_t(\mathbf{w}) - \widehat{\ell}_t(\mathbf{w}^\star)\right]\right] \\
&\quad= \mathbb{E}\left[\sum_{t=1}^{T}\left(\sum_{\mathbf{w}\in V} x_t(\mathbf{w})\ell_t(\mathbf{w}) - \ell_t(\mathbf{w}^\star)\right)\right].
\end{aligned}$$

The first equality follows from linearity of expectation and the tower property: for each t, we average first over the current draw conditional on $\mathcal{F}_{t-1}$ and then over all histories up to round $t-1$. This yields the same expectation as averaging over the full trajectory through round T, since the round-t term is determined by the history up to round t and is unaffected by subsequent draws. No independence across rounds is required. The second equality follows from conditional unbiasedness and the fact that x_t is fixed conditional on $\mathcal{F}_{t-1}$.

Similarly, applying the tower property and (D.5) yields

$$\mathbb{E}\left[\sum_{t=1}^{T}\sum_{\mathbf{w}\in V} x_t(\mathbf{w})\widehat{\ell}_t(\mathbf{w})^2\right] = \sum_{t=1}^{T}\mathbb{E}\left[\mathbb{E}_t\left[\sum_{\mathbf{w}\in V} x_t(\mathbf{w})\widehat{\ell}_t(\mathbf{w})^2\right]\right] \le \frac{dT}{1-\gamma}.$$

Taking expectations in (D.11) and substituting these relations gives

$$\mathbb{E}\left[\sum_{t=1}^{T}\left(\sum_{\mathbf{w}\in V} x_t(\mathbf{w})\ell_t(\mathbf{w}) - \ell_t(\mathbf{w}^\star)\right)\right] \le \frac{\log|V|}{\eta} + \frac{\eta dT}{1-\gamma}. \tag{D.17}$$

Step 2: Account for actual sampling and exploration. Since $\mathbf{w}(t) \sim p_t$, $\mathbb{E}_t[\ell_t(\mathbf{w}(t))] = \sum_{\mathbf{w}\in V} p_t(\mathbf{w})\ell_t(\mathbf{w})$. Furthermore,

$$\begin{aligned}
\sum_{\mathbf{w}\in V}(p_t(\mathbf{w}) - x_t(\mathbf{w}))\ell_t(\mathbf{w}) &= \gamma\left(\sum_{\mathbf{w}\in V}\mu(\mathbf{w})\ell_t(\mathbf{w}) - \sum_{\mathbf{w}\in V} x_t(\mathbf{w})\ell_t(\mathbf{w})\right) \\
&\le 2\gamma.
\end{aligned} \tag{D.18}$$

The last inequality uses $\ell_t(\mathbf{w}) \in [-1, 1]$ and the fact that μ, x_t are probability distributions.

To use the bound already obtained for x_t, we add and subtract its average loss in each round:

$$\begin{aligned}
\mathbb{E}_t[\ell_t(\mathbf{w}(t))] - \ell_t(\mathbf{w}^\star) &= \sum_{\mathbf{w}\in V} p_t(\mathbf{w})\ell_t(\mathbf{w}) - \ell_t(\mathbf{w}^\star) \\
&= \sum_{\mathbf{w}\in V} x_t(\mathbf{w})\ell_t(\mathbf{w}) - \ell_t(\mathbf{w}^\star) + \sum_{\mathbf{w}\in V}(p_t(\mathbf{w}) - x_t(\mathbf{w}))\ell_t(\mathbf{w}) \\
&\le \sum_{\mathbf{w}\in V} x_t(\mathbf{w})\ell_t(\mathbf{w}) - \ell_t(\mathbf{w}^\star) + 2\gamma,
\end{aligned}$$

where the last inequality follows from (D.18). Thus, the actual sampling distribution increases the conditional expected loss by at most 2γ per round compared with x_t.

Taking full expectations and summing over t, the tower property gives

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T}\ell_t(\mathbf{w}(t)) - \sum_{t=1}^{T}\ell_t(\mathbf{w}^\star)\right] \\
&= \sum_{t=1}^{T}\mathbb{E}[\mathbb{E}_t[\ell_t(\mathbf{w}(t))] - \ell_t(\mathbf{w}^\star)] \\
&\le \mathbb{E}\left[\sum_{t=1}^{T}\left(\sum_{\mathbf{w}\in V} x_t(\mathbf{w})\ell_t(\mathbf{w}) - \ell_t(\mathbf{w}^\star)\right)\right] + 2\gamma T \\
&\le \frac{\log|V|}{\eta} + \frac{\eta d T}{1-\gamma} + 2\gamma T, \qquad \text{(D.19)}
\end{aligned}$$

where the final inequality uses (D.17).

By normalization and exact finite reduction,

$$\mathbb{E}[R(T)] = CF\,\mathbb{E}\left[\sum_{t=1}^{T}\ell_t(\mathbf{w}(t)) - \sum_{t=1}^{T}\ell_t(\mathbf{w}^\star)\right]. \qquad \text{(D.20)}$$

This proves general regret bound in theorem 2.

Step 3: Choose parameters and retain the original scale. Under (5.8), $\gamma = \eta d^2 \le 1/2$. Therefore

$$\frac{\mathbb{E}[R(T)]}{CF} \le \frac{\log|V|}{\eta} + 2\eta d T + 2\eta d^2 T \le \frac{\log|V|}{\eta} + 4\eta d^2 T. \qquad \text{(D.21)}$$

If $T \ge d^2\log|V|$, then $\eta = (2d)^{-1}\sqrt{\log|V|/T}$. Both terms on the right of (D.21) equal $2d\sqrt{T\log|V|}$, proving $\mathbb{E}[R(T)] \le 4CFd\sqrt{T\log|V|}$. If $T < d^2\log|V|$, the bound $|r_t(\mathbf{w})| \le CF$ gives instead

$$\mathbb{E}[R(T)] \le 2CFT < 2CFd\sqrt{T\log|V|},$$

which is stronger than the stated bound. Finally, $d \le KF$ and $|V| \le \binom{N+K\binom{N}{2}}{N-1}$ give the last inequality.

Their cumulative payoff equals the continuous optimum by Lemma 1. □

D.5 Optimized exploration with a fixed exploration rate

We retain the payoff coordinates $\mathbf{c}(\mathbf{w})$, the loss estimator, and the exponential-weights update of Algorithm 2. Following the exploration principle of Bubeck et al. (2012), we optimize the exploration distribution in the existing payoff coordinates. Let $\Delta(V)$ denote the probability distributions on V, and set

$$\begin{aligned}
\mu &\in \arg\max_{\mu'\in\Delta(V)} \log\det\left(\sum_{\mathbf{w}\in V}\mu'(\mathbf{w})\mathbf{c}(\mathbf{w})\mathbf{c}(\mathbf{w})^\top\right), \\
S_\mu &= \sum_{\mathbf{w}\in V}\mu(\mathbf{w})\mathbf{c}(\mathbf{w})\mathbf{c}(\mathbf{w})^\top.
\end{aligned} \qquad \text{(D.22)}$$

The objective is assigned the value $-\infty$ at singular matrices. This distribution is computed once from the known payoff features.

Proposition 3 (Optimized exploration with a fixed exploration rate)**.** *Retain the model, fixed oblivious sequence, and finite candidate set of Theorem 2, with $d \ge 1$ and $|V| \ge 2$. Fix $\eta > 0$ with $\eta d \le 1/2$, and set $\gamma = \eta d$. Modify the exploration rule of Algorithm 2 by using*

$$p_t = (1-\gamma)x_t + \gamma\mu, \qquad S_t = \sum_{\mathbf{w}\in V} p_t(\mathbf{w})\mathbf{c}(\mathbf{w})\mathbf{c}(\mathbf{w})^\top. \qquad \text{(D.23)}$$

Both η and γ remain fixed throughout the run. With the original estimator and the unchanged update (D.10)*, the modified algorithm satisfies*

$$\begin{aligned}\mathbb{E}[R(T)] &\le CF\left(\frac{\log|V|}{\eta} + \frac{\eta d T}{1-\gamma} + 2\gamma T\right) \\ &\le CF\left(\frac{\log|V|}{\eta} + 4\eta d T\right).\end{aligned} \tag{D.24}$$

In particular, choosing

$$\eta = \min\left\{\frac{1}{2d}, \frac{1}{2}\sqrt{\frac{\log|V|}{dT}}\right\}, \qquad \gamma = \eta d \tag{D.25}$$

gives, for every $T \ge 1$,

$$\mathbb{E}[R(T)] \le 4CF\sqrt{dT\log|V|}. \tag{D.26}$$

Proof. The proof reuses the unbiasedness, second-moment, and exponential-weights calculations above. The key additional step is to obtain a sharper bound on the loss estimates from the optimized exploration distribution.

Since $\mathbf{c}(\mathbf{w}_i) = e_i$, the uniform distribution on the basis strategies has second moment I_d/d. The determinant is continuous on the compact probability simplex, and its maximum is therefore attained and strictly positive. Thus (D.22) has a maximizer with $S_\mu \succ 0$.

For any $\mathbf{w} \in V$, moving an ε fraction of the exploration distribution to $\mathbf{w}$, with $0 \le \varepsilon < 1$, gives a feasible distribution. Since μ maximizes the log-determinant, the corresponding right derivative is nonpositive. By Jacobi's formula,

$$\begin{aligned}0 &\ge \frac{\mathrm{d}}{\mathrm{d}\varepsilon}\log\det\big((1-\varepsilon)S_\mu + \varepsilon\mathbf{c}(\mathbf{w})\mathbf{c}(\mathbf{w})^\top\big)\Big|_{\varepsilon=0^+} \\ &= \operatorname{tr}\big[S_\mu^{-1}\big(\mathbf{c}(\mathbf{w})\mathbf{c}(\mathbf{w})^\top - S_\mu\big)\big] \\ &= \mathbf{c}(\mathbf{w})^\top S_\mu^{-1}\mathbf{c}(\mathbf{w}) - d.\end{aligned} \tag{D.27}$$

The last equality follows from $\operatorname{tr}(Azz^\top) = z^\top Az$, $S_\mu^{-1}S_\mu = I_d$, and $\operatorname{tr}(I_d) = d$. Thus, $\mathbf{c}(\mathbf{w})^\top S_\mu^{-1}\mathbf{c}(\mathbf{w}) \le d$ for every $\mathbf{w} \in V$.

Since $0 < \gamma = \eta d \le 1/2$, the sampling distribution gives

$$S_t = (1-\gamma)\sum_{\mathbf{v}\in V} x_t(\mathbf{v})\mathbf{c}(\mathbf{v})\mathbf{c}(\mathbf{v})^\top + \gamma S_\mu \succeq \gamma S_\mu \succ 0.$$

Inversion reverses the order of positive-definite matrices, so

$$S_t^{-1} \preceq \frac{1}{\gamma}S_\mu^{-1}.$$

Consequently, for every $\mathbf{w} \in V$,

$$\eta\mathbf{c}(\mathbf{w})^\top S_t^{-1}\mathbf{c}(\mathbf{w}) \le \frac{\eta}{\gamma}\mathbf{c}(\mathbf{w})^\top S_\mu^{-1}\mathbf{c}(\mathbf{w}) \le \frac{\eta d}{\gamma} = 1. \tag{D.28}$$

Using $|\ell_t(\mathbf{w}(t))| \le 1$ and the Cauchy–Schwarz inequality, we obtain

$$\begin{aligned}|\widehat{\ell}_t(\mathbf{w})| &= \big|\mathbf{c}(\mathbf{w})^\top S_t^{-1}\mathbf{c}(\mathbf{w}(t))\ell_t(\mathbf{w}(t))\big| \\ &\le \sqrt{\mathbf{c}(\mathbf{w})^\top S_t^{-1}\mathbf{c}(\mathbf{w})}\sqrt{\mathbf{c}(\mathbf{w}(t))^\top S_t^{-1}\mathbf{c}(\mathbf{w}(t))} \le \frac{1}{\eta}.\end{aligned} \tag{D.29}$$

Hence Lemma 4 applies without modification.

Conditional on $\mathcal{F}_{t-1}$, the distributions x_t, p_t and the matrix S_t are fixed. The identities in (D.6) and (D.8)–(D.9) use only the actual sampling distribution and its invertible second-moment matrix. They therefore remain valid and give

$$\begin{aligned}\mathbb{E}_t[\widehat{\ell}_t(\mathbf{w})] &= \ell_t(\mathbf{w}), \\ \mathbb{E}_t\left[\sum_{\mathbf{w}\in V} x_t(\mathbf{w})\widehat{\ell}_t(\mathbf{w})^2\right] &\le \frac{d}{1-\gamma}.\end{aligned} \tag{D.30}$$

The second inequality also uses $p_t(\mathbf{w}) \geq (1-\gamma)x_t(\mathbf{w})$.

Fix $\mathbf{w}^\star \in \arg\min_{\mathbf{w}\in V} \sum_{t=1}^{T} \ell_t(\mathbf{w})$. It is independent of the learner's randomization because the attacker sequence is fixed. Taking expectations in (D.11) and using (D.30) yields

$$\mathbb{E}\sum_{t=1}^{T}\left(\sum_{\mathbf{w}\in V} x_t(\mathbf{w})\ell_t(\mathbf{w}) - \ell_t(\mathbf{w}^\star)\right) \leq \frac{\log|V|}{\eta} + \frac{\eta d T}{1-\gamma}. \tag{D.31}$$

Since $p_t = (1-\gamma)x_t + \gamma\mu$ and $\ell_t(\mathbf{w}) \in [-1,1]$, the conditional exploration cost satisfies

$$\begin{aligned}\sum_{\mathbf{w}\in V}(p_t(\mathbf{w}) - x_t(\mathbf{w}))\ell_t(\mathbf{w}) &= \gamma\left(\sum_{\mathbf{w}\in V}\mu(\mathbf{w})\ell_t(\mathbf{w}) - \sum_{\mathbf{w}\in V} x_t(\mathbf{w})\ell_t(\mathbf{w})\right)\\ &\leq 2\gamma.\end{aligned}$$

Using $\mathbf{w}(t) \sim p_t$, the scale conversion (D.20), and (D.31), we obtain

$$\begin{aligned}\frac{\mathbb{E}[R(T)]}{CF} &= \mathbb{E}\sum_{t=1}^{T}\left(\sum_{\mathbf{w}\in V} p_t(\mathbf{w})\ell_t(\mathbf{w}) - \ell_t(\mathbf{w}^\star)\right)\\ &\leq \mathbb{E}\sum_{t=1}^{T}\left(\sum_{\mathbf{w}\in V} x_t(\mathbf{w})\ell_t(\mathbf{w}) - \ell_t(\mathbf{w}^\star)\right) + 2\gamma T\\ &\leq \frac{\log|V|}{\eta} + \frac{\eta d T}{1-\gamma} + 2\gamma T.\end{aligned}$$

Multiplying by CF proves the first inequality in (D.24). Since $\gamma = \eta d \leq 1/2$,

$$\frac{\eta d T}{1-\gamma} + 2\gamma T \leq 2\eta d T + 2\eta d T = 4\eta d T,$$

which proves its second inequality.

For $T \geq d\log|V|$, the choice in (D.25) gives

$$\frac{\log|V|}{\eta} = 4\eta d T = 2\sqrt{dT\log|V|}.$$

For $T < d\log|V|$, we instead use $|r_t(\mathbf{w})| \leq CF$ to obtain

$$\mathbb{E}[R(T)] \leq 2CFT < 2CF\sqrt{dT\log|V|}.$$

This proves (D.26) for every $T \geq 1$. □

E Proof of the horizon lower bound

Proof of Proposition 2. We construct a fixed RSSG instance for which every online algorithm incurs expected regret of order $\sqrt{T}$ on some oblivious sequence.

Step 1: A two-target RSSG instance. Consider

$$N = 2, \qquad K = 1, \qquad F = 2,$$

and let the defender strategy be

$$\mathbf{w} = (w_1, w_2), \qquad w_1 + w_2 = 1, \qquad w_1, w_2 \geq 0.$$

Choose the defender utilities as

$$\lambda_1^r = -\frac{1}{4}, \qquad \rho_1^r = \frac{1}{4}, \qquad \lambda_2^r = -\frac{1}{2}, \qquad \rho_2^r = 1.$$

There is only one attacker type, whose utilities are

$$\lambda_1^1 = \lambda_2^1 = 1, \qquad \rho_1^1 = \rho_2^1 = 0.$$

All of these utilities satisfy the ranges assumed in the model.

The attacker's utilities for the two targets are

$$u_a^1(E_1, \mathbf{w}) = 1 - w_1, \qquad u_a^1(E_2, \mathbf{w}) = 1 - w_2 = w_1.$$

Hence, when the attack intensity is one, the attacker chooses target 1 if $w_1 < 1/2$ and target 2 if $w_1 > 1/2$.

At $w_1 = w_2 = 1/2$, the attacker is indifferent between the two targets. The corresponding defender utilities are

$$U(\mathbf{w}, 1) = -\frac{1}{4} + \frac{1}{2}w_1 = 0$$

and

$$U(\mathbf{w}, 2) = -\frac{1}{2}(1 - w_2) + w_2 = \frac{1}{4}.$$

Therefore, the optimistic tie-breaking rule selects target 2.

It follows that the defender payoff against an intensity-one attacker is

$$f_1^1(\mathbf{w}) = \begin{cases} -\frac{1}{4} + \frac{1}{2}w_1, & 0 \le w_1 < \frac{1}{2}, \\ 1 - \frac{3}{2}w_1, & \frac{1}{2} \le w_1 \le 1. \end{cases} \tag{E.1}$$

When the attack intensity is two, both targets are attacked, and therefore

$$\begin{aligned} f_2^1(\mathbf{w}) &= \left(-\frac{1}{4} + \frac{1}{2}w_1\right) + \left(1 - \frac{3}{2}w_1\right) \\ &= \frac{3}{4} - w_1. \end{aligned} \tag{E.2}$$

Step 2: Two fixed defender strategies. Consider

$$\mathbf{w}^A = (0, 1), \qquad \mathbf{w}^B = \left(\frac{1}{2}, \frac{1}{2}\right).$$

Equations (E.1)–(E.2) give

	intensity 1	intensity 2
$\mathbf{w}^A$	$-\frac{1}{4}$	$\frac{3}{4}$
$\mathbf{w}^B$	$\frac{1}{4}$	$\frac{1}{4}$.

(E.3)

We next verify that no other feasible defender strategy has larger expected payoff when the two attack intensities occur with equal probability. From (E.1)–(E.2),

$$\frac{1}{2}f_1^1(\mathbf{w}) + \frac{1}{2}f_2^1(\mathbf{w}) = \begin{cases} \frac{1}{4} - \frac{1}{4}w_1, & 0 \le w_1 < \frac{1}{2}, \\ \frac{7}{8} - \frac{5}{4}w_1, & \frac{1}{2} \le w_1 \le 1. \end{cases} \tag{E.4}$$

Since the first expression is at most $1/4$, and the second satisfies

$$\frac{7}{8} - \frac{5}{4}w_1 \le \frac{7}{8} - \frac{5}{8} = \frac{1}{4}$$

for $w_1 \ge 1/2$, we obtain

$$\frac{1}{2}f_1^1(\mathbf{w}) + \frac{1}{2}f_2^1(\mathbf{w}) \le \frac{1}{4} \qquad \text{for every } \mathbf{w} \in W. \tag{E.5}$$

Step 3: A random oblivious sequence. We first consider a random sequence of follower configurations. Before the game starts, independently for every round t, draw the attack intensity l_t uniformly from $\{1, 2\}$.

There are exactly C followers in every round, all with the same realized intensity. Thus,

$$M(t) = \begin{cases} (C, 0), & l_t = 1, \\ (0, C), & l_t = 2. \end{cases}$$

This satisfies

$$\sum_{k=1}^{K} \sum_{\ell=1}^{F} m_{k\ell}(t) = C$$

in every round.

Fix an arbitrary, possibly randomized, learning algorithm. Conditional on the entire history before round t and on its selected strategy $\mathbf{w}(t)$, the current intensity l_t is still uniform over $\{1, 2\}$. Hence

$$\begin{aligned} \mathbb{E}\left[r_t(\mathbf{w}(t)) \,|\, \text{history and } \mathbf{w}(t)\right] &= \frac{C}{2} f_1^1(\mathbf{w}(t)) + \frac{C}{2} f_2^1(\mathbf{w}(t)) \\ &\leq \frac{C}{4}, \end{aligned} \tag{E.6}$$

where the inequality follows from (E.5).

Taking expectations and summing over t gives

$$\mathbb{E}\left[\sum_{t=1}^{T} r_t(\mathbf{w}(t))\right] \leq \frac{CT}{4}. \tag{E.7}$$

Notice that this argument does not depend on what feedback is revealed after each round. In particular, (E.7) remains valid even under full-information feedback.

Step 4: Payoff of the hindsight benchmark. Define

$$X_t = \begin{cases} -1, & l_t = 1, \\ +1, & l_t = 2, \end{cases} \qquad S_T = \sum_{t=1}^{T} X_t.$$

The variables $X_1, \ldots, X_T$ are independent random variables.

From (E.3),

$$r_t(\mathbf{w}^A) = \frac{C}{4} + \frac{C}{2} X_t,$$

while

$$r_t(\mathbf{w}^B) = \frac{C}{4}$$

for both possible attack intensities. Therefore,

$$\sum_{t=1}^{T} r_t(\mathbf{w}^A) = \frac{CT}{4} + \frac{C}{2} S_T, \tag{E.8}$$

and

$$\sum_{t=1}^{T} r_t(\mathbf{w}^B) = \frac{CT}{4}. \tag{E.9}$$

Since the hindsight benchmark maximizes over all feasible defender strategies, it is at least as large as the better of these two:

$$\begin{aligned} \max_{\mathbf{w} \in W} \sum_{t=1}^{T} r_t(\mathbf{w}) &\geq \max\left\{ \sum_{t=1}^{T} r_t(\mathbf{w}^A), \sum_{t=1}^{T} r_t(\mathbf{w}^B) \right\} \\ &= \frac{CT}{4} + \frac{C}{2} \max\{S_T, 0\}. \end{aligned} \tag{E.10}$$

Combining (E.7) and (E.10) yields

$$\begin{aligned}\mathbb{E}[R(T)] &\geq \frac{C}{2}\mathbb{E}[\max\{S_T, 0\}] \\ &= \frac{C}{4}\mathbb{E}|S_T|, \end{aligned} \tag{E.11}$$

where the last equality follows from the symmetry of S_T.

Step 5: A lower bound for the random walk. Since the X_t are independent with

$$\mathbb{E}[X_t] = 0, \qquad X_t^2 = 1,$$

we have

$$\mathbb{E}[S_T^2] = T. \tag{E.12}$$

For the fourth moment, all terms containing an odd power of some X_t have zero expectation. Hence

$$\begin{aligned}\mathbb{E}[S_T^4] &= \sum_{t=1}^{T}\mathbb{E}[X_t^4] + 6\sum_{1\leq s<t\leq T}\mathbb{E}[X_s^2 X_t^2] \\ &= T + 6\binom{T}{2} \\ &= 3T^2 - 2T \leq 3T^2. \end{aligned} \tag{E.13}$$

Applying Hölder's inequality,

$$\begin{aligned}\mathbb{E}[S_T^2] &= \mathbb{E}\left[|S_T|^{2/3}|S_T|^{4/3}\right] \\ &\leq \left(\mathbb{E}|S_T|\right)^{2/3}\left(\mathbb{E}|S_T|^4\right)^{1/3}. \end{aligned} \tag{E.14}$$

Rearranging and using (E.12) and (E.13),

$$\begin{aligned}\mathbb{E}|S_T| &\geq \frac{\left(\mathbb{E}[S_T^2]\right)^{3/2}}{\left(\mathbb{E}[S_T^4]\right)^{1/2}} \\ &\geq \frac{T^{3/2}}{\sqrt{3}T} = \sqrt{\frac{T}{3}}. \end{aligned} \tag{E.15}$$

Substituting (E.15) into (E.11) gives

$$\mathbb{E}[R(T)] \geq \frac{C}{4\sqrt{3}}\sqrt{T}. \tag{E.16}$$

Step 6: From the random construction to an oblivious deterministic sequence. The expectation in (E.16) is currently taken over both the random intensity sequence and the learner's internal randomization. Equivalently,

$$\mathbb{E}_{l_1,\ldots,l_T}\left[\mathbb{E}_{\text{alg}}\left[R(T) \mid l_1, \ldots, l_T\right]\right] \geq \frac{C}{4\sqrt{3}}\sqrt{T}.$$

Therefore, there exists at least one deterministic realization $(l_1, \ldots, l_T) \in \{1, 2\}^T$ such that

$$\mathbb{E}_{\text{alg}}\left[R(T) \mid l_1, \ldots, l_T\right] \geq \frac{C}{4\sqrt{3}}\sqrt{T}.$$

Fix this realization before the game starts. The corresponding sequence $\{M(t)\}_{t=1}^T$ is therefore oblivious, and after fixing it the only remaining expectation is over the learner's internal randomization.

This proves

$$\mathbb{E}[R(T)] \geq \frac{C}{4\sqrt{3}}\sqrt{T},$$

and completes the proof. □

F Exact Computation for Attacker Sequences

To solve this generalized problem, we extend our previous single-level MILP reduction in 3.3. Specifically, we generalize components of the MILP: the decision variables, the linearization of bilinear terms, and the encoding of the attacker best-response conditions via Big-M constraints, to account for multiple attacker types and intensities. The resulting MILP computes the defender's exact optimal allocation given the empirical count matrix $\hat{\mathbf{M}}(t)$:

$$\max_{\mathbf{w},\{\mathbf{z}^k_{1:F}\}_{k=1}^K,\{\mathbf{h}^k_{1:F}\}_{k=1}^K,\{\tau^k_{1:F}\}_{k=1}^K} \quad \sum_{k=1}^{K}\sum_{l=1}^{F} \hat{m}_{kl}\cdot R_l^k(\mathbf{w},\mathbf{z}_l^k) \tag{F.1a}$$

$$\text{s.t.} \quad \sum_{\mathcal{E}_j\in\mathcal{E}^-} w_j = 1, \quad w_j \geq 0, \quad \forall k\in\{1,\dots,K\}, \forall l\in\{1,\dots,F\}, \forall \mathcal{E}_j\in\mathcal{E}^- : \tag{F.1b}$$

$$R_l^k(\mathbf{w},\mathbf{z}_l^k) = \sum_{\mathcal{E}_j\in\mathcal{E}^-} \left[\lambda_j^r h_{j,l}^k - \lambda_j^r z_{j,l}^k + \rho_j^r z_{j,l}^k\right], \quad z_{j,l}^k \leq w_j, \quad z_{j,l}^k \geq w_j + h_{j,l}^k - 1, \tag{F.1c}$$

$$\sum_{\mathcal{E}_j\in\mathcal{E}^-} h_{j,l}^k = l, \quad \mathbf{h}_l^k \geq \mathbf{h}_{l-1}^k \quad (l\geq 2), \tag{F.1d}$$

$$\lambda_j^k(1-w_j) + \rho_j^k w_j \geq \tau_l^k - B(1-h_{j,l}^k), \quad \lambda_j^k(1-w_j) + \rho_j^k w_j \leq \tau_l^k + B h_{j,l}^k, \tag{F.1e}$$

$$z_{j,l}^k \in [0,1], \quad z_{j,l}^k \leq h_{j,l}^k, \quad h_{j,l}^k \in \{0,1\}, \quad -1 \leq \tau_l^k \leq 1. \tag{F.1f}$$

G Empirical Efficiency of the Oracle

We compare the wall-clock runtime of a single MILP oracle call with exhaustive LP enumeration. For each configuration, we generate five independent random game instances. Within each repetition, both methods receive the same game parameters and oracle objective weights. The enumeration baseline considers every ordered list of F distinct targets for each of the K attacker types, yielding $\big(N!/(N-F)!\big)^K$ candidate LPs. Timing covers the complete oracle calls, including model construction and optimization. Enumeration time includes all candidate LPs.

All 50 paired runs completed successfully and passed the objective agreement check

$$|v_{\mathrm{MILP}} - v_{\mathrm{LP}}| \leq 10^{-6}\max\{1, |v_{\mathrm{MILP}}|, |v_{\mathrm{LP}}|\},$$

where v_{MILP} and v_{LP} denote the objective values returned by the two methods.

Table 2 shows that enumeration is faster for the smallest instance and has a similar runtime at $(N,K,F) = (4,2,2)$. Its cost then increases substantially with the number of candidate LPs. In the fixed-$(K,F) = (3,2)$ series, increasing N from 4 to 9 raises enumeration time from 2.757 to 596.190 seconds, while the MILP mean remains between 0.333 and 0.436 seconds. At the largest tested configuration, $(N,K,F) = (9,3,2)$, the MILP oracle achieves approximately a $1{,}439\times$ speedup. These results provide empirical evidence that the MILP implementation is efficient relative to explicit LP enumeration over the tested range.

The experiments were implemented in MATLAB R2024b Update 1 (version 24.2.0.2740171), with YALMIP R20230622 used to formulate the MILPs and Gurobi 9.0.0 used to solve both the MILPs and the LPs in the enumeration baseline.

Table 2: Runtime comparison between the MILP oracle and exhaustive LP enumeration. Times are reported as mean $\pm$ sample standard deviation, in seconds, over five independent repetitions. Speedup is the ratio of the enumeration mean to the MILP mean; values above one favor the MILP oracle.

N	K	F	# LPs	MILP (s)	Enumeration (s)	Speedup
3	2	2	36	0.240 ± 0.015	0.063 ± 0.004	$0.26\times$
4	2	2	144	0.245 ± 0.014	0.230 ± 0.013	$0.94\times$
5	2	2	400	0.271 ± 0.026	0.640 ± 0.024	$2.36\times$
4	2	3	576	0.325 ± 0.033	0.910 ± 0.016	$2.80\times$
4	3	2	1,728	0.333 ± 0.026	2.757 ± 0.031	$8.29\times$
5	3	2	8,000	0.366 ± 0.020	12.807 ± 0.072	$35.00\times$
6	3	2	27,000	0.392 ± 0.014	43.401 ± 0.472	$110.71\times$
7	3	2	74,088	0.402 ± 0.038	119.350 ± 0.914	$296.80\times$
8	3	2	175,616	0.436 ± 0.047	281.760 ± 3.452	$646.27\times$
9	3	2	373,248	0.414 ± 0.043	596.190 ± 7.053	$1439.40\times$

H STOCHASTIC MULTI-TYPE ATTACKER DISTRIBUTIONS

Multi-Type Attack Distribution in Stackelberg Games We consider a Stackelberg setting involving a single leader and $n \geq 1$ followers. The leader has $N = |\mathcal{E}| \geq 2$ actions and selects a mixed strategy $\mathbf{w} \in \mathcal{W}$ over them. over these actions. Each follower has a finite action set $\mathcal{E}^-$ with cardinality $|\mathcal{E}^-| = N$, and the joint action of all followers is denoted by $\mathbf{v} = (v_1, \ldots, v_n)$.

In each round, every follower i is associated with a private type $\alpha_i \in [K]$ and an attack intensity $l_i \in \{1, \ldots, F\}$. We denote a follower of type α_k with attack intensity l as θ_l^k. The joint type profile of all followers is defined as $\boldsymbol{\theta} = (\theta_{l_1}^{\alpha_1}, \ldots, \theta_{l_n}^{\alpha_n}) \in \Theta^n$, which is drawn from a bayesian setting distribution $\mathcal{D}$, i.e., $\boldsymbol{\theta} \sim \mathcal{D}$. In this context, we consider a general type distribution in which follower types and attack intensities may be arbitrarily correlated. Let $m_{kl} = \sum_{i=1}^{n} \mathbb{I}(\theta_{l_i}^{\alpha_i} = \theta_l^k)$, which denotes the number of attackers of type α_k with attack intensity l at a given time step. The utility function associated with type θ_l^k is defined as:

$$f_l^k(\mathbf{w}) = \sum_{\mathcal{E}_j \in \mathbf{b}_k(\mathbf{w})} \lambda_j^r(1 - w_j) + \rho_j^r w_j, \tag{H.1a}$$

$$\text{s.t.} \quad \sum_{\mathcal{E}_j \in \mathcal{E}^-} w_j = 1, \quad \mathbf{b}_k(\mathbf{w}) = \arg\max_{\mathcal{E}_f^k \cap \mathcal{E}^-} \sum_{\mathcal{E}_j \in \mathcal{E}_f^k \cap \mathcal{E}^-} u_a^k(\mathcal{E}_j, \mathbf{w}), \quad |\mathbf{b}_k(\mathbf{w})| = l, \tag{H.1b}$$

where $\mathcal{E}_f^k$ is the targets set selected by the attacker α_k and $u_a^k(\mathcal{E}_j, \mathbf{w}) = \lambda_j^k(1 - w_j) + \rho_j^k w_j$. Let $\mathbf{M} \in \mathbb{R}^{K \times F}$ be the matrix with entrie $(\mathbf{M})_{k,l} = m_{kl}$, where m_{kl} denotes the number of attackers of type k with intensity l. Similarly, define $\mathbf{F}(\mathbf{w}) \in \mathbb{R}^{K \times F}$ with entries $(\mathbf{F}(\mathbf{w}))_{k,l} = f_l^k(\mathbf{w})$. The total utility can then be compactly expressed as the sum of element-wise products $\sum_{k=1}^{K} \sum_{l=1}^{F} \big(\mathbf{M} \odot \mathbf{F}(\mathbf{w})\big)_{kl}$, where $\odot$ denotes the Hadamard (entrywise) product.

Based on this, the corresponding single-level MILP for the multi-type setting is given as follows:

$$\max_{\mathbf{w}, \{\mathbf{z}_{1:F}^k\}_{k=1}^K, \{\mathbf{h}_{1:F}^k\}_{k=1}^K, \{\tau_{1:F}^k\}_{k=1}^K} \quad \sum_{k=1}^{K} \sum_{l=1}^{F} m_{kl} \cdot R_l^k(\mathbf{w}, \mathbf{z}_l^k) \tag{H.2a}$$

$$\text{s.t.} \quad \sum_{\mathcal{E}_j \in \mathcal{E}^-} w_j = 1, \quad w_j \geq 0, \quad \forall k \in \{1, \ldots, K\}, \forall l \in \{1, \ldots, F\}, \forall \mathcal{E}_j \in \mathcal{E}^- : \tag{H.2b}$$

$$R_l^k(\mathbf{w}, \mathbf{z}_l^k) = \sum_{\mathcal{E}_j \in \mathcal{E}^-} \left[\lambda_j^r h_{j,l}^k - \lambda_j^r z_{j,l}^k + \rho_j^r z_{j,l}^k\right], \quad z_{j,l}^k \leq w_j, \; z_{j,l}^k \geq w_j + h_{j,l}^k - 1, \tag{H.2c}$$

$$\sum_{\mathcal{E}_j \in \mathcal{E}^-} h_{j,l}^k = l, \quad \sum_{k=1}^{K} \sum_{l=1}^{F} m_{kl} = n, \quad \mathbf{h}_l^k \geq \mathbf{h}_{l-1}^k \; (l \geq 2), \tag{H.2d}$$

$$\lambda_j^k(1 - w_j) + \rho_j^k w_j \geq \tau_l^k - B(1 - h_{j,l}^k), \quad \lambda_j^k(1 - w_j) + \rho_j^k w_j \leq \tau_l^k + B h_{j,l}^k, \tag{H.2e}$$

$$z_{j,l}^k \in [0, 1], \quad h_{j,l}^k \in \{0, 1\}, \quad z_{j,l}^k \leq h_{j,l}^k, \quad -1 \leq \tau_l^k \leq 1. \tag{H.2f}$$

Refer to Figure 1b. In our experiments, the attacker and defender utility matrices are defined as follows. The (i, j)-th entry of each matrix represents the utility of attacker j attacking target i.

Attacker Utilities

$$U_{\text{attacker}}^u = \begin{bmatrix} 0.55 & 0.79 & 0.70 & 0.85 & 0.27 \\ 0.73 & 0.47 & 0.98 & 0.40 & 0.33 \\ 0.80 & 0.46 & 0.65 & 0.41 & 0.59 \\ 0.32 & 0.07 & 0.64 & 0.15 & 0.34 \\ 0.89 & 0.31 & 0.84 & 0.05 & 0.20 \end{bmatrix},$$

$$U_{\text{attacker}}^c = \begin{bmatrix} -0.88 & -0.02 & -0.76 & -0.53 & -0.50 \\ -0.51 & -0.42 & -0.34 & -0.09 & -0.33 \\ -0.20 & -0.65 & -0.79 & -0.84 & -0.86 \\ -0.61 & -0.21 & -0.39 & -0.15 & -0.62 \\ -0.39 & -0.16 & -0.92 & -0.36 & -0.97 \end{bmatrix}.$$

Defender Utilities

$$U^c_{\text{defender}} = [0.99 \quad 0.42 \quad 0.47 \quad 0.89 \quad 0.83]\,,$$

$$U^u_{\text{defender}} = [-0.16 \quad -0.97 \quad -0.40 \quad -0.26 \quad -0.84]\,.$$

Here, U^u_{attacker} and U^c_{attacker} correspond to the attacker's utilities when the target is unprotected and protected, respectively. Similarly, U^u_{defender} and U^c_{defender} represent the defender's utilities under the same conditions.

I ADVERSARIAL SEQUENCES WITH ONE ATTACKER TYPE

In this section, we illustrate a concrete example for the single-follower case shown in fig 2a, where there is only one attacker type active in each round. The adversary can design a sequence of attack intensities over multiple rounds to systematically deceive the defender. By carefully constructing this sequence, the attacker ensures that the defender, who chooses strategies based only on historical empirical frequencies, repeatedly selects suboptimal strategies. This leads to a linear accumulation of regret over time, demonstrating how adversarial attack intensity can exploit the defender's reliance on past observations.

The adversary designs a repeated SSG to induce linear regret of the defender in Table 3, 4.

Table 3: Defender's utility matrix

Protected Target	**Attack Target 1**	**Attack Target 2**
Target 1	$\frac{1}{4}$	-$\frac{1}{4}$
Target 2	$-\frac{1}{4}$	1

Table 4: Attacker's utility matrix

Protected Target	**Attack Target 1**	**Attack Target 2**
Target 1	0	1
Target 2	$\frac{1}{2}$	0

We can split the feasible set into two regions with different follower's response when attack intensity $l = 1$:

$$\text{region 1}: \; 1 - w_1 \geq 2w_1, \qquad \text{region 2}: \; 1 - w_1 \leq 2w_1,$$

with boundary point $\left(\frac{1}{3}, \frac{2}{3}\right)$. In region 1, the attacker will best respond to attack target for higher expected utility. In region 2, the best response is attacking target 2.

For attack intensity $l = 1$, we have

$$\begin{aligned} \mathbf{w} \in \text{region 1}: \quad & R_1(\mathbf{w}) = -\frac{1}{4}(1 - w_1) + \frac{1}{4}w_1 = -\frac{1}{4} + \frac{1}{2}w_1, \\ \mathbf{w} \in \text{region 2}: \quad & R_1(\mathbf{w}) = -\frac{1}{4}(1 - w_2) + w_2 = -\frac{1}{4} + \frac{5}{4}w_2. \end{aligned}$$

For attack intensity $l = 2$, the note writes the combined term:

$$R_2(\mathbf{w}) = \frac{1}{4}\big(-(1 - w_1) + w_1\big) + \frac{1}{4}\big(-(1 - w_2) + 4w_2\big).$$

Let q_1 and q_2 be the frequencies of intensities $l = 1$ and $l = 2$, respectively. The expected utility of the defender is

$$\begin{aligned}
\mathbf{w} \in \text{region 1}: \quad \langle \mathbf{f}(\mathbf{w}), \mathbf{q} \rangle &= R_1(\mathbf{w})q_1 + R_2(\mathbf{w})q_2 \\
&= \frac{1}{4}\big(-(1-w_1) + w_1\big)q_1 \\
&\quad + \frac{1}{4}\big(-(1-w_1) + w_1 - (1-w_2) + 4w_2\big)q_2, \\
\mathbf{w} \in \text{region 2}: \quad \langle \mathbf{f}(\mathbf{w}), \mathbf{q} \rangle &= R_1(\mathbf{w})q_1 + R_2(\mathbf{w})q_2 \\
&= \frac{1}{4}\big(-(1-w_2) + 4w_2\big)q_1 \\
&\quad + \frac{1}{4}\big(-(1-w_1) + w_1 - (1-w_2) + 4w_2\big)q_2.
\end{aligned}$$

Taking the polytope/extreme-point argument into each region, we get

$$\begin{aligned}
\text{region 1}: \quad & \mathbf{w} = (0,1), && \langle \mathbf{f}(\mathbf{w}), \mathbf{q} \rangle = -\frac{1}{4}q_1 + \frac{3}{4}q_2, \\
& \mathbf{w} = \left(\frac{1}{3}, \frac{2}{3}\right), && \langle \mathbf{f}(\mathbf{w}), \mathbf{q} \rangle = -\frac{1}{12}q_1 + \frac{1}{2}q_2; \\
\text{region 2}: \quad & \mathbf{w} = \left(\frac{1}{3}, \frac{2}{3}\right), && \langle \mathbf{f}(\mathbf{w}), \mathbf{q} \rangle = \frac{7}{12}q_1 + \frac{1}{2}q_2, \\
& \mathbf{w} = (1,0), && \langle \mathbf{f}(\mathbf{w}), \mathbf{q} \rangle = -\frac{1}{4}q_1.
\end{aligned}$$

By evaluating the vertices of each region (extreme points) with tie-breaking in favor of defender, we obtain the set of candidate strategies for the defender that may be optimal in each region:

$$\text{A}: \ \mathbf{w} = (0,1) \text{ (region 1)}, \qquad \text{B}: \ \mathbf{w} = \left(\frac{1}{3}, \frac{2}{3}\right) \text{ (region 2)}.$$

By comparing A and B, we have

$$-q_1 + 3q_2 > \frac{7}{3}q_1 + 2q_2 \iff q_2 > \frac{10}{3}q_1,$$

which means A is better than B when $q_2 > \frac{10}{3}q_1$; otherwise, $q_2 < \frac{10}{3}q_1$, B is better than A.

The above formulation assumes that the attack intensity is known in advance. In practice, however, its distribution is unknown and only empirical frequencies from historical data are available. We therefore base our analysis on these empirical estimates.

Adversarial sequence for q_1, q_2

Assume that the defender in round t can only observe attack intensities up to round $t-1$. Accordingly, it selects a strategy that is optimal with respect to the empirical estimates $(\hat{q}_1(t-1), \hat{q}_2(t-1))$. At each round t, the attacker selects an intensity $l \in \{1, 2\}$. The defender observes only the history up to round $t-1$, forms empirical frequency estimates accordingly

$$\hat{q}_1(t-1) = \frac{1}{t-1}\sum_{s=1}^{t-1} \mathbb{I}\{l_s = 1\}, \qquad \hat{q}_2(t-1) = \frac{1}{t-1}\sum_{s=1}^{t-1} \mathbb{I}\{l_s = 2\},$$

and plays the strategy optimal for $(\hat{q}_1(t-1), \hat{q}_2(t-1))$.

Recall that

$$\langle \mathbf{f}(\mathbf{w}_A), \mathbf{q} \rangle = -\frac{1}{4}q_1 + \frac{3}{4}q_2, \qquad \langle \mathbf{f}(\mathbf{w}_B), \mathbf{q} \rangle = \frac{7}{12}q_1 + \frac{1}{2}q_2,$$

so we get

$$\langle \mathbf{f}(\mathbf{w}_A), \hat{\mathbf{q}} \rangle - \langle \mathbf{f}(\mathbf{w}_B), \hat{\mathbf{q}} \rangle = \frac{1}{4}\hat{q}_2 - \frac{10}{12}\hat{q}_1.$$

Hence A is optimal iff $\hat{q}_2 > \frac{10}{3}\hat{q}_1$, and B is optimal iff $\hat{q}_2 < \frac{10}{3}\hat{q}_1$.

Now use a 13-round cycle of actual attacks:

$$I_{13m+n} = \begin{cases} 1, & n \in \{2, 6, 10\}, \\ 2, & \text{otherwise}, \end{cases} \qquad m, n \in N,\ n \in [0, 12]$$

n	1	2	3	4	5	6	7
l_t	2	1	2	2	2	1	2
$\hat{q}(t-1)$	$(0,0)$	$(0,1)$	$(\frac{1}{2},\frac{1}{2})$	$(\frac{1}{3},\frac{2}{3})$	$(\frac{1}{4},\frac{3}{4})$	$(\frac{1}{5},\frac{4}{5})$	$(\frac{1}{3},\frac{2}{3})$
$\hat{q}(t)$	$(0,1)$	$(\frac{1}{2},\frac{1}{2})$	$(\frac{1}{3},\frac{2}{3})$	$(\frac{1}{4},\frac{3}{4})$	$(\frac{1}{5},\frac{4}{5})$	$(\frac{1}{3},\frac{2}{3})$	$(\frac{2}{7},\frac{5}{7})$
$\mathbf{w}^*_{t-1}$	$\sim$	A	B	B	B	A	B
$\mathbf{w}^*_t$	A	B	B	B	A	B	B

n	8	9	10	11	12	13
l_t	2	2	1	2	2	2
$\hat{q}(t-1)$	$(\frac{2}{7},\frac{5}{7})$	$(\frac{1}{4},\frac{3}{4})$	$(\frac{2}{9},\frac{7}{9})$	$(\frac{3}{10},\frac{7}{10})$	$(\frac{3}{11},\frac{8}{11})$	$(\frac{4}{12},\frac{8}{12})$
$\hat{q}(t)$	$(\frac{1}{4},\frac{3}{4})$	$(\frac{2}{9},\frac{7}{9})$	$(\frac{3}{10},\frac{7}{10})$	$(\frac{3}{11},\frac{8}{11})$	$(\frac{4}{12},\frac{8}{12})$	$(\frac{3}{13},\frac{10}{13})$
$\mathbf{w}^*_{t-1}$	B	B	A	B	B	B
$\mathbf{w}^*_t$	B	A	B	B	B	$\sim$

The defender chooses strategy $\mathbf{w}_t$ in round t according to $(\hat{q}_1(t-1), \hat{q}_2(t-1))$. For this period-13 construction, $q_2 - \frac{10}{3}q_1 = 0$ repeatedly oscillates around 0. Strategy $\mathbf{w}^*_t$ is selected by optimizing $\langle \mathbf{f}(\mathbf{w}), \hat{\mathbf{q}}_t \rangle$. The strategy $\mathbf{w}_t = \mathbf{w}^*_{t-1}$ is selected by follow the leader algorithm (not follow the perturbed leader), $\mathbf{w}^*_t$ is selected by being the leader algorithm. Deterministic strategies $\mathbf{w}_t$ based on empirical frequencies ($\hat{\mathbf{q}}$) expose the defenders to exploitation by adaptive adversaries that are capable to infer protection patterns. Thus, adversaries may anticipate protection patterns and attack uncovered targets. Hence, in every 13-round cycle, there exist five rounds in which the defender picks the wrong strategy between A and B.

Therefore, after m full cycles ($T = 13m$), cumulative regret satisfies $R(T) = \sum_{t=1}^{T} r_t \geq \frac{9}{4}m$.

We note that in Figures 2a and 3b, Baseline 2 performs better than Algorithm 1. This is because, in the single-attacker-type setting, Baseline 2 trivially converges to a fixed mixed strategy, which coincides with strategy B. In contrast, our adversarial sequence is specifically constructed to mislead the follow-the-leader (FTL) algorithm by creating two strategies, A and B, whose cumulative payoffs remain nearly identical over most rounds. As a result, FTL is repeatedly induced to select the suboptimal strategy, whereas Baseline 2 consistently plays strategy B. Moreover, during the $13m$ rounds, both A and B achieve the same payoff as the hindsight optimal strategy, and regret is incurred only during the remaining n rounds where the two strategies differ. Therefore, the superior empirical performance of Baseline 2 in this setting is largely a consequence of this particular construction and is essentially trivial.

J Adversarial sequences with multiple attacker types

In this section, we illustrate a concrete example for the single-follower case shown in fig 2b, where in each round there is only one attacker, but the attacker may be of different types and attack intensities. The adversary can design a sequence of attacker types and attack intensities over multiple rounds to systematically deceive the defender. By carefully constructing this sequence, the attacker ensures that the defender, who chooses strategies based only on historical empirical frequencies, repeatedly selects suboptimal strategies. This leads to a linear accumulation of regret over time, demonstrating how an adversarially designed sequence of attacker types and attack intensities can exploit the defender's reliance on past observations.

The adversary designs a repeated SSG to induce linear regret of the defender in Table 5, 6.

We can partition the defender's strategy space into three regions according to the attackers' best responses:

- **Region 1**: Both attacker α_1 and α_2 choose attack target 1 as their best response.

Table 5: Defender's utility matrix

Protected Target	Attack Target 1	Attack Target 2
Target 1	$\frac{1}{4}$	$-\frac{1}{4}$
Target 2	$-\frac{1}{4}$	1

Table 6: Attacker α_1's utility matrix

Protected Target	Attack Target 1	Attack Target 2
Target 1	0	1
Target 2	$\frac{1}{2}$	0

Table 7: Attacker α_2's utility matrix

Protected Target	Attack Target 1	Attack Target 2
Target 1	0	$\frac{1}{2}$
Target 2	1	0

- **Region 2**: Attacker 1 chooses target 2 and attacker 2 chooses target 1.
- **Region 3**: Both attacker types choose target 2 as their best response .

Specifically, attacker 1's decision boundary is at $(\frac{1}{3}, \frac{2}{3})$ and attacker 2's decision boundary is at $(\frac{2}{3}, \frac{1}{3})$, which naturally divides the strategy space into three regions of distinct best-response combinations.

On each region, the defender's expected utility is a linear function of $\mathbf{w}$, and by the polytope/extreme-point argument, the candidate optimal strategies lie at the vertices of each region. Hence, the possible candidate strategies are:

$$\begin{aligned}
&\text{Region 1:} \quad \mathbf{w} = (0,1), \quad \mathbf{w} = \left(\frac{1}{3}, \frac{2}{3}\right), \\
&\text{Region 2:} \quad \mathbf{w} = \left(\frac{1}{3}, \frac{2}{3}\right), \quad \mathbf{w} = \left(\frac{2}{3}, \frac{1}{3}\right), \\
&\text{Region 3:} \quad \mathbf{w} = \left(\frac{2}{3}, \frac{1}{3}\right), \quad \mathbf{w} = (1,0).
\end{aligned}$$

By evaluating the vertices of each region (extreme points) with tie-breaking in favor of defender, we obtain the set of candidate strategies for the defender that may be optimal in each region.

The candidate strategies are:

$$\begin{aligned}
\text{A}: \ \mathbf{w} &= (0,1) \text{ (region 1)}, \\
\text{B}: \ \mathbf{w} &= \left(\frac{1}{3}, \frac{2}{3}\right) \text{ (region 2)}. \\
\text{C}: \ \mathbf{w} &= (\frac{2}{3}, \frac{1}{3}) \text{ (region 3)}.
\end{aligned}$$

Given distribution $\mathbf{q} = (q_{1,1}, q_{1,2}, q_{2,1}, q_{2,2}) \in \Delta_4$, where q_{kl} represents probability of attacker type α_k and attack intensity l. The utility function is written as:

$$\begin{aligned}
\text{A}: \ \langle \mathbf{f}(\mathbf{w}_A), \mathbf{q} \rangle &= -\frac{1}{4} q_{1,1} - \frac{1}{4} q_{2,1} + \frac{3}{4} q_{1,2} + \frac{3}{4} q_{2,2}, \\
\text{B}: \ \langle \mathbf{f}(\mathbf{w}_B), \mathbf{q} \rangle &= \frac{7}{12} q_{1,1} - \frac{1}{12} q_{2,1} + \frac{1}{2} q_{1,2} + \frac{1}{2} q_{2,2},
\end{aligned}$$

$$\mathrm{C}: \ \langle \mathbf{f}(\mathbf{w}_C), \mathbf{q} \rangle = \frac{1}{6}q_{1,1} + \frac{1}{6}q_{2,1} + \frac{1}{4}q_{1,2} + \frac{1}{4}q_{2,2},$$

WLOG, we focus on two strategies A and B, which are induced as optimal strategies for the attacker under an estimated sequence $\hat{\mathbf{q}}$. Here, $\hat{\mathbf{q}}$ is adversarially constructed so as to steer the attacker into selecting A and B as optimal responses. The equation is written as

$$-\frac{1}{4}q_{1,1} - \frac{1}{4}q_{2,1} + \frac{3}{4}q_{1,2} + \frac{3}{4}q_{2,2} = \frac{7}{12}q_{1,1} - \frac{1}{12}q_{2,1} + \frac{1}{2}q_{1,2} + \frac{1}{2}q_{2,2},$$

which characterizes the indifference condition between strategies A and B. It suffices that the empirical distribution constructed by the attacker oscillates across the two sides of this equality, so that A and B are alternately optimal. Then we have $H(\mathbf{q}) = \frac{13}{3}q_{1,1} + \frac{5}{3}q_{2,1} - 1$, where defender will choose strategy A if $H(\mathbf{q}) \leq 0$ and choose strategy B otherwise.

The foregoing outlines the decision-making criteria under the assumption that the attack intensity is known; however, since it is impossible to know its distribution in advance—and typically, one can only observe historical empirical frequencies—we will proceed to illustrate the process using these empirical frequencies.

Adversarial sequence Defender at round t can only observe attack intensities up to round $t-1$, and therefore uses at round t the strategy that is optimal under $(\hat{q}_{1,1}(t-1), \hat{q}_{1,2}(t-1), \hat{q}_{2,1}(t-1), \hat{q}_{2,2}(t-1))$. At each round, the defender only observes history up to $t-1$, forms empirical frequencies

$$\hat{q}_{k,l}(t-1) = \frac{1}{t-1}\sum_{s=1}^{t-1} \mathbb{I}\{\alpha_s = \alpha_k, l_s = l\},$$

and at round t plays the strategy optimal for $\hat{\mathbf{q}}(t-1)$.

Now use a 6-round cycle of actual attacks:

$$I_{6m+n} = \begin{cases} \alpha = \alpha_1, l = 1, & n = 4, \\ \alpha = \alpha_2, l = 1, & n = 1, \\ \alpha = \alpha_1 \text{ or } \alpha_2, l = 2, & \text{otherwise}, \end{cases}$$

where $m, n \in N, n \in [1, 6]$.

n	1	2	3	4	5	6
α_t, l_t	α_2,1	α_2,2	α_2,2	α_1,1	α_2,2	α_1,2
$m \cdot \mathbf{H}(\hat{\mathbf{q}}(t-1))$	0	$\frac{2}{3}$	$-\frac{1}{6}$	$-\frac{4}{9}$	$\frac{1}{2}$	$\frac{1}{5}$
$m \cdot \mathbf{H}(\hat{\mathbf{q}}(t))$	$\frac{2}{3}$	$-\frac{1}{6}$	$-\frac{4}{9}$	$\frac{1}{2}$	$\frac{1}{5}$	0
$\mathbf{w}^*_{t-1}$	$\sim$	B	A	A	B	B
$\mathbf{w}^*_t$	B	A	A	B	B	$\sim$

The defender chooses strategy $\mathbf{w}_t$ in round t according to $\hat{\mathbf{q}}(t-1))$. For this construction, $H(\hat{\mathbf{q}})$ repeatedly oscillates around 0. Strategy $\mathbf{w}^*_t$ is selected by optimizing $\langle \mathbf{f}(\mathbf{w}), \hat{\mathbf{q}}_t \rangle$. Strategy $\mathbf{w}_t = \mathbf{w}^*_{t-1}$ is selected by follow the leader algorithm (not follow the perturbed leader), $\mathbf{w}^*_t$ is selected by being the leader algorithm. Deterministic strategies $\mathbf{w}_t = \mathbf{w}^*_{t-1}$ based on empirical frequencies ($\hat{\mathbf{q}}$) expose the defenders to exploitation by adaptive adversaries that are capable to infer protection patterns. Thus, adversaries may anticipate protection patterns and attack uncovered targets.

Hence, in every 6-round cycle, there exist two rounds in which the defender picks the wrong strategy between A and B. Therefore, after m full cycles ($T = 6m$), cumulative regret satisfies $R(T) = \sum_{t=1}^{T} r_t \geq \frac{7}{12}m$.

K ADVERSARIAL SEQUENCES WITH MULTIPLE FOLLOWERS AND TYPES

In this section, we consider a multiple-follower setting shown in fig 1a, where a fixed number of attackers appear in each round. While the number of followers remains constant, their types and

attack intensities may vary across rounds. We construct an adversarial sequence over attacker types and attack intensities to systematically mislead the defender. By carefully coordinating the joint behavior of multiple followers, the adversary ensures that the empirical distribution observed by the defender induces alternating optimal strategies, causing the defender to repeatedly select suboptimal actions. As a result, the defender incurs linear regret over time.

Consider the multiple follower scenario (two follower). The utility matrices are referred to Table 5, 6, 7. Given vector $\mathbf{q} = (q_{1,1}, q_{1,2}, q_{2,1}, q_{2,2})$, where q_{kl} represents frequency of attacker type α_k and attack intensity l and $\sum_{k,l} q_{kl} = 2$.

By evaluating the vertices of each region (extreme points) with tie-breaking in favor of defender, we obtain the set of candidate strategies for the defender that may be optimal in each region. The candidate strategies are:

$$\text{A}: \ \mathbf{w} = (0, 1) \text{ (region 1)},$$

$$\text{B}: \ \mathbf{w} = \left(\frac{1}{3}, \frac{2}{3}\right) \text{ (region 2)}.$$

$$\text{C}: \ \mathbf{w} = (\frac{2}{3}, \frac{1}{3}) \text{ (region 3)}.$$

Given distribution $\mathbf{q} = (q_{1,1}, q_{1,2}, q_{2,1}, q_{2,2})$, with $\sum_{k,l} q_{kl} = 2$. The utility function is written as:

$$\text{A}: \ \langle \mathbf{f}(\mathbf{w}_A), \mathbf{q} \rangle = -\frac{1}{4}q_{1,1} - \frac{1}{4}q_{2,1} + \frac{3}{4}q_{1,2} + \frac{3}{4}q_{2,2},$$

$$\text{B}: \ \langle \mathbf{f}(\mathbf{w}_B), \mathbf{q} \rangle = \frac{7}{12}q_{1,1} - \frac{1}{12}q_{2,1} + \frac{1}{2}q_{1,2} + \frac{1}{2}q_{2,2},$$

$$\text{C}: \ \langle \mathbf{f}(\mathbf{w}_C), \mathbf{q} \rangle = \frac{1}{6}q_{1,1} + \frac{1}{6}q_{2,1} + \frac{1}{4}q_{1,2} + \frac{1}{4}q_{2,2},$$

We focus on two strategies A and B, which are induced as optimal strategies for the attacker under an estimated sequence $\hat{\mathbf{q}}$. Here, $\hat{\mathbf{q}}$ is adversarially constructed so as to steer the attacker into selecting A and B as optimal responses. The equation is written as

$$-\frac{1}{4}q_{1,1} - \frac{1}{4}q_{2,1} + \frac{3}{4}q_{1,2} + \frac{3}{4}q_{2,2} = \frac{7}{12}q_{1,1} - \frac{1}{12}q_{2,1} + \frac{1}{2}q_{1,2} + \frac{1}{2}q_{2,2},$$

which characterizes the indifference condition between strategies A and B. It suffices that the empirical distribution constructed by the attacker oscillates across the two sides of this equality, so that A and B are alternately optimal. Then we have $H(\mathbf{q}) = \frac{13}{3}q_{1,1} + \frac{5}{3}q_{2,1} - 2$, where defender will choose strategy A if $H(\mathbf{q}) \leq 0$ and choose strategy B otherwise.

The foregoing outlines the decision-making criteria under the assumption that the attack intensity is known; however, since it is impossible to know its distribution in advance—and typically, one can only observe historical empirical frequencies—we will proceed to illustrate the process using these empirical frequencies.

Adversarial sequence Defender at round t can only observe attack intensities up to round $t-1$, and therefore uses at round t the strategy that is optimal under $(\hat{q}_{1,1}(t-1), \hat{q}_{1,2}(t-1), \hat{q}_{2,1}(t-1), \hat{q}_{2,2}(t-1))$. At each round, the defender only observes history up to $t-1$, forms empirical frequencies

$$\hat{q}_{k,l}(t-1) = \frac{1}{t-1} \sum_{s=1}^{t-1} \sum_{c=1}^{2} \mathbb{I}\{\alpha_{s,c} = \alpha_k, l_{s,c} = l\},$$

,where c denotes the follower c. At round t plays the strategy optimal for $\hat{\mathbf{q}}(t-1)$.

Now use a 3-round cycle of actual attacks:

where $m, n \in N, n \in [1, 3]$.

n	1	2	3
follower 1	α_2,2	α_1,1	α_2,2
follower 2	α_1,2	α_2,2	α_2,1
$m \cdot \mathbf{H}(\hat{\mathbf{q}}(t-1))$	0	-2	$\frac{1}{6}$
$m \cdot \mathbf{H}(\hat{\mathbf{q}}(t))$	-2	$\frac{1}{6}$	0
$\mathbf{w}_{t-1}^*$	$\sim$	A	B
$\mathbf{w}_t^*$	A	B	$\sim$

The defender chooses strategy $\mathbf{w}_t$ in round t according to $\hat{\mathbf{q}}(t-1))$. For this construction, $H(\hat{\mathbf{q}})$ repeatedly oscillates around 0. Strategy $\mathbf{w}_t^*$ is selected by optimizing $\langle \mathbf{f}(\mathbf{w}), \hat{\mathbf{q}}_t \rangle$. Strategy $\mathbf{w}_t = \mathbf{w}_{t-1}^*$ is selected by follow the leader algorithm (not follow the perturbed leader), $\mathbf{w}_t^*$ is selected by being the leader algorithm. Deterministic strategies $\mathbf{w}_t = \mathbf{w}_{t-1}^*$ based on empirical frequencies ($\hat{\mathbf{q}}$) expose the defenders to exploitation by adaptive adversaries that are capable to infer protection patterns. Thus, adversaries may anticipate protection patterns and attack uncovered targets.

Hence, in every 3-round cycle, there exist one rounds in which the defender picks the wrong strategy between A and B. Therefore, after m full cycles ($T = 3m$), cumulative regret satisfies $R(T) = \sum_{t=1}^{T} r_t \geq \frac{m}{12}$.

L Experiments with a time-varying follower population

In this section, we consider a setting with a time-varying number of followers under full information feedback shown in fig 3b. In each round, multiple attackers may appear, but the number of active followers can vary (e.g., no-shows). We construct an adversarial sequence over both the number of active followers and their attack intensities to systematically mislead the defender. By carefully designing this sequence, the adversary ensures that the empirical frequency estimates based on historical observations repeatedly induce incorrect strategy choices. As a result, the defender is forced to alternate between suboptimal decisions and incurs linear regret over time. Additionally, in partial information feedback scenario, only aggregate attack effect are observable without identifying individual attack intensities or follower numbers.

Consider the multiple follower scenario (two follower). The utility matrices are referred to Table 3, 4. Given vector $\mathbf{q} = (q_{1,1}, q_{1,2})$, where q_l represents frequency of attacker's with attack intensity l.

By evaluating the vertices of each region (extreme points) with tie-breaking in favor of defender, we obtain the set of candidate strategies for the defender that may be optimal in each region:

$$\text{A}: \ \mathbf{w} = (0,1) \text{ (region 1)}, \qquad \text{B}: \ \mathbf{w} = \left(\frac{1}{3}, \frac{2}{3}\right) \text{ (region 2)}.$$

By comparing A and B, we have

$$-q_1 + 3q_2 > \frac{7}{3}q_1 + 2q_2 \iff q_2 > \frac{10}{3}q_1,$$

which means A is better than B when $q_2 > \frac{10}{3}q_1$; otherwise, $q_2 < \frac{10}{3}q_1$, B is better than A.

The above formulation assumes that the attack intensity frequency is known in advance. In practice, however, its frequency is unknown and only empirical frequencies from historical data are available. We therefore base our analysis on these empirical estimates.

Adversarial sequence for q_1, q_2

Assume that the defender in round t can only observe attack intensities up to round $t-1$. Accordingly, it selects a strategy that is optimal with respect to the empirical estimates $(\hat{q}_1(t-1), \hat{q}_2(t-1))$. At each round t, the attacker selects an intensity $l \in \{1, 2\}$. The defender observes only the history up to round $t-1$, forms empirical frequency estimates accordingly

$$\hat{q}_1(t-1) = \frac{1}{t-1} \sum_{s=1}^{t-1} \sum_{c=1}^{2} \mathbb{I}\{l_{s,c} = 1\}, \qquad \hat{q}_2(t-1) = \frac{1}{t-1} \sum_{s=1}^{t-1} \sum_{c=1}^{2} \mathbb{I}\{l_{s,c} = 2\},$$

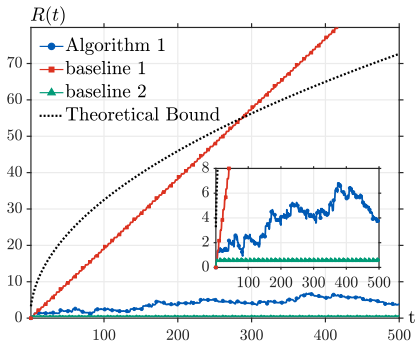
and plays the strategy optimal for $(\hat{q}_1(t-1), \hat{q}_2(t-1))$.

Now use a 7-round cycle of actual attacker sequences:

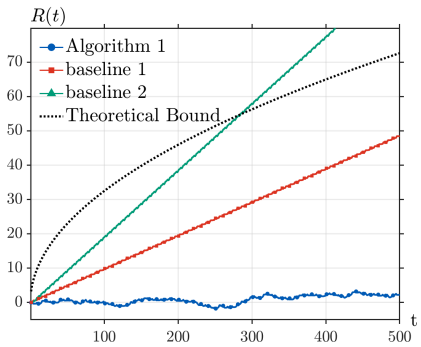
n	1	2	3	4	5	6	7
follower 1	l=2	l=1	l=2	l=1	l=2	l=1	l=2
follower 2	l=2	no show	l=2	l=2	l=2	l=2	l=2
$\mathbf{w}_{t-1}$	$\sim$	A	B	A	B	A	B
$\mathbf{w}_t$	A	B	A	B	A	B	$\sim$

The defender chooses strategy $\mathbf{w}_t$ in round t according to $(\hat{q}_1(t-1), \hat{q}_2(t-1))$. For this period-7 construction, $q_2 - \frac{10}{3}q_1 = 0$ repeatedly oscillates around 0. Strategy $\mathbf{w}_t^*$ is selected by optimizing $\langle \mathbf{f}(\mathbf{w}), \hat{\mathbf{q}}_t \rangle$. The strategy $\mathbf{w}_t = \mathbf{w}_{t-1}^*$ is selected by follow the leader algorithm (not follow the perturbed leader), $\mathbf{w}_t^*$ is selected by being the leader algorithm. Deterministic strategies $\mathbf{w}_t$ based on empirical frequencies $(\hat{\mathbf{q}})$ expose the defenders to exploitation by adaptive adversaries that are capable to infer protection patterns. Thus, adversaries may anticipate protection patterns and attack uncovered targets. Hence, in every 7-round cycle, there exist five rounds in which the defender picks the wrong strategy between A and B.

Therefore, after m full cycles ($T = 7m$), cumulative regret satisfies $R(T) = \sum_{t=1}^{T} r_t \geq \frac{3}{2}m$.

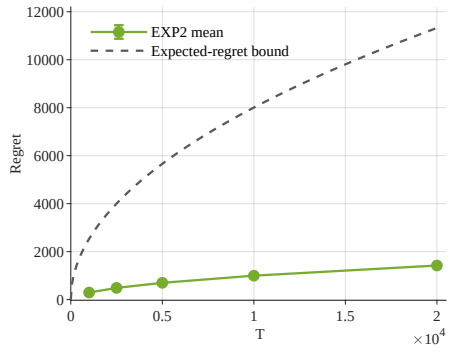


(a) Single attacker type with adversarial attack intensity sequence

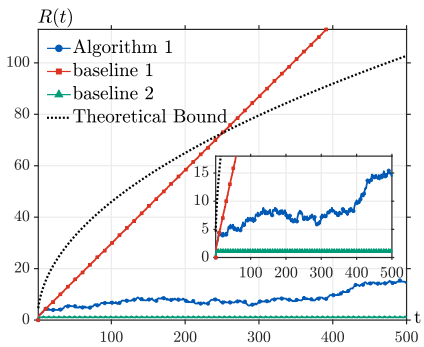


(b) Multiple attacker type with adversarial attack intensity sequence

Figure 2: Regret comparison under different adversarial settings.



(a) barycentric-spanner EXP2



(b) Time-varying follower number with adversarial attack intensity sequence

Figure 3: Retained results: (a) barycentric-spanner EXP2; (b) full-information FPL with a time-varying follower population.

M AGGREGATE-FEEDBACK LEARNING WITH TWO FOLLOWERS

We consider two targets, two attacker types, and exactly two followers per round, so that $N = K = F = C = 2$. The defender chooses $\mathbf{w} \in [0,1]^2$ with $w_1 + w_2 = 1$. Each follower attacks one or two targets, and best-response ties favor the defender. The utility parameters are

$$\boldsymbol{\lambda}^r = (-1,-1), \qquad \boldsymbol{\rho}^r = (1,1), \qquad U_a^u = \begin{bmatrix} 1/4 & 1 \\ 1 & 1/4 \end{bmatrix}, \qquad U_a^c = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}.$$

Rows index targets and columns index attacker types. The candidate set is

$$V = \{(0,1), (1/5, 4/5), (4/5, 1/5), (1,0)\}.$$

The intensity-two payoffs are zero for every strategy, and the payoff vectors have effective dimension $d = 2$. We use the first two candidates as the basis strategies.

Experimental setting. Every 250 rounds, the type probabilities alternate between $(0.85, 0.15)$ and $(0.75, 0.25)$, while the intensity probabilities alternate between $(0.9, 0.1)$ and $(0.8, 0.2)$. Types and intensities are sampled independently for each follower and round, conditional on the regime. The complete attacker sequence is generated before learning. Both regimes have the same unique expected-optimal strategy, $(1/5, 4/5)$, with payoff gaps of 1.512 and 0.960 to the second-best candidate. This stable optimum makes the instance suitable for illustrating how the algorithm learns from aggregate feedback.

Since $\log|V|/T$ can be small, we use a larger constant multiplier in the learning rate while retaining its $T^{-1/2}$ scaling and enforcing $\gamma = d^2\eta \le 1/2$, which changes only the constant in the $\mathcal{O}(\sqrt{T})$ regret bound.

Parameter choice. We use the original EXP2 update and normalization $\ell_t(\mathbf{w}) = -r_t(\mathbf{w})/(CF)$. For each total horizon T, choose

$$\eta = \min\left\{\frac{1}{2d^2}, \frac{2}{d}\sqrt{\frac{\log|V|}{T}}\right\}, \qquad \gamma = d^2\eta. \tag{M.1}$$

Both parameters remain fixed throughout each run. The learning-rate coefficient was selected in a separate preliminary comparison of one, two, and four times the original recommended coefficient.

The general regret bound allows any parameters satisfying $\eta d^2 \le \gamma$ and $0 < \gamma < 1$. Our choice satisfies $\eta d^2 = \gamma \le 1/2$, so the original proof applies:

$$\begin{aligned}\mathbb{E}[R(T)] &\le CF\left[\frac{\log|V|}{\eta} + \frac{\eta d T}{1-\gamma} + 2\gamma T\right] \\ &\le CF\left[\frac{\log|V|}{\eta} + 4\eta d^2 T\right].\end{aligned}$$

For $T \ge 16d^2\log|V|$, substitution gives

$$\mathbb{E}[R(T)] \le \frac{17}{2}CFd\sqrt{T\log|V|} = 68\sqrt{T\log 4}. \tag{M.2}$$

For smaller T, the same bound follows from $R(T) \le 2CFT \le 8CFd\sqrt{T\log|V|}$. Thus, the larger learning rate preserves the square-root dependence on T, although it increases the bound's constant.

Results. We use $T \in \{1000, 2500, 5000, 10000, 20000\}$, with ten independently generated attacker sequences and 32 algorithm repetitions per sequence. For each horizon, the algorithm starts from uniform weights and uses the fixed parameters in (M.1). Regret is evaluated against the best fixed candidate in hindsight; only the selected strategy's aggregate payoff is used for learning.

For $T = 10{,}000$, the parameters are $\eta = 0.01177410$ and $\gamma = 0.04709640$. The mean probability of selecting $(1/5, 4/5)$ reaches 0.9489 at round 1,000 and 0.9763 at round 2,500. The final mean cumulative regret is 999.59, with standard error 2.52 across the ten sequence-level means.

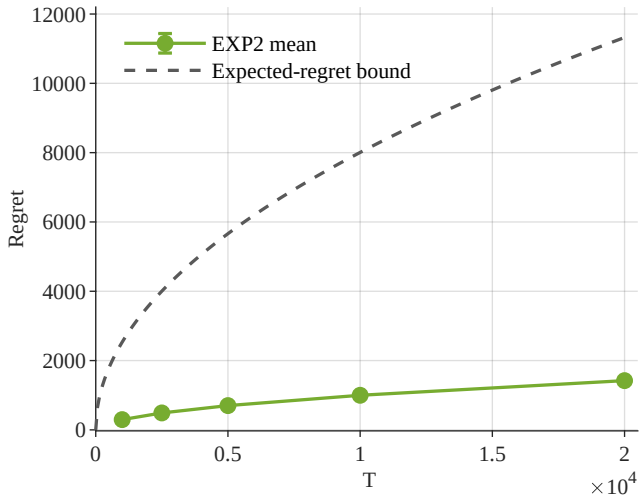


Figure 4: Aggregate-feedback EXP2 with fixed parameters..