When to Reset Your Keys: Optimal Timing of Security Updates via Learning

Zizhan Zheng
Department of Computer Science
Tulane University

Ness B. Shroff
Dept. of ECE and CSE
The Ohio State University

Prasant Mohapatra
Department of Computer Science
University of California, Davis

Abstract

Cybersecurity is increasingly threatened by advanced and persistent attacks. As these attacks are often designed to disable a system (or a critical resource, e.g., a user account) repeatedly, it is crucial for the defender to keep updating its security measures to strike a balance between the risk of being compromised and the cost of security updates. Moreover, these decisions often need to be made with limited and delayed feedback due to the stealthy nature of advanced attacks. In addition to targeted attacks, such an optimal timing policy under incomplete information has broad applications in cybersecurity. Examples include key rotation, password change, application of patches, and virtual machine refreshing. However, rigorous studies of optimal timing are rare. Further, existing solutions typically rely on a pre-defined attack model that is known to the defender, which is often not the case in practice. In this work, we make an initial effort towards achieving optimal timing of security updates in the face of unknown stealthy attacks. We consider a variant of the influential FlipIt game model with asymmetric feedback and unknown attack time distribution, which provides a general model to consecutive security updates. The defender's problem is then modeled as a time associative bandit problem with dependent arms. We derive upper confidence bound based learning policies that achieve low regret compared with optimal periodic defense strategies that can only be derived when attack time distributions are known.

Introduction

Malicious attacks are constantly evolving to inflict increasing levels of damage on the nation's infrastructure systems, cooperate IT systems, and our digital lives. For example, the Advanced Persistent Threat (APT) has become a major concern to cybersecurity in the past few years. APT attacks exhibit two distinguishing behavior patterns (van Dijk et al. 2013) that make them extremely difficult to defend using traditional techniques. First, these attacks are often funded well and persistent. They attack a target system (or a critical resource) *periodically* with the goal to compromise it *completely* e.g., by stealing full cryptography keys. Second, the attacks can be highly adaptive. In particular, they often act *covertly*, e.g., by operating in a "low-and-slow" fashion (Bowers et al. 2014), to avoid immediate detection and obtain long-term advantages.

From the defender's perspective, an effective way to thwart continuous and stealthy attacks is to update its security measures periodically to strike a balance between the risk of being compromised and the cost of updates. The primary challenge, however, is that such decisions must often be made with limited and delayed feedback because of the covert nature of the attacker. In addition to thwarting targeted attacks, such an *optimal timing* problem with incomplete information is crucial in various cybersecurity scenarios, e.g., key rotation (van Dijk et al. 2013), password changes (Tan and Xia 2016), application of patches (Beattie et al. 2002), and virtual machine refreshing (Juels et al. 2016). For example, Facebook receives approximately 600,000 "compromised logins" from impostors every day (Barnett 2011). An efficient approach to stop these attacks is to ask users to update their passwords when the risk of attack is high.

Although time-related tactical security choices have been studied since the cold war era (Blackwell 1949), rigorous study of timing decisions in the face of continuous and stealthy attacks is relatively new. In 2012, in response to an APT attack on it, the RSA lab proposed the FlitIt game, which was one of the first models to study timing decisions under stealthy takeovers. The FlipIt game model abstracts out details about concrete attack and defense operations by focusing on the stealthy and persistent nature of players. The basic model considers two players, each of whom can "flip" the state of a system periodically at any time with a cost. A player only learns the system state when she moves herself. The payoff of a player is defined as the fraction of time when the resource is under its control less the total cost incurred.

The FlipIt game captures the stealthy behavior of players in an elegant way by allowing various types of feedback structures. In the basic model where neither player gets any feedback during the game and each move flips the state of the resource instantaneously, it is known that periodic strategies with random starting phases form a pair of best response strategies (van Dijk et al. 2013). As a variant of the basic model, an asymmetric setting is studied in (Laszka, Johnson, and Grossklags 2013) where the defender gets no feedback during the game while the attacker obtains immediate feedback after each defense but incurs a random attack time to

take over the resource. In this setting, it is shown in (Laszka, Johnson, and Grossklags 2013) that periodic defense and immediate attack (or no attack) form a pair of best response strategies. However, little is known beyond these two cases. In particular, designing adaptive defense strategies with partial feedback remains an open problem.

Although the FlipIt game provides a proper framework to understand the strategic behavior of stealthy takeover, it relies on detailed prior knowledge about the attacker. In particular, it requires parameters such as the amount of time needed to compromise a resource and the unit cost of each attack (or their distributions) to be fixed and known to the defender so that the equilibrium solution can be derived. These parameters limit the scope of the attack model, which, however, can be hard to verify before the game starts. To address this fundamental limitation, we propose to study online learning algorithms that make minimum assumptions about the attacker and learn an optimal defense strategy from the limited feedback obtained during the game. Given the advances in big data analytics and their applications in cybersecurity, it is feasible for the defender to obtain partial feedback even under stealthy attacks. Such a learning approach makes it possible to derive adaptive and robust defense strategies against *unknown attacks* where the type of the attacker is derived from a fixed but unknown distribution, as well as the more challenging *dynamic attacks* where the type of the attacker can arbitrarily vary over time.

In this work, we make a first effort towards achieving optimal timing of security updates in the face of unknown stealthy attacks. We consider a variant of FlipIt game with asymmetric feedback similar to (Laszka, Johnson, and Grossklags 2013), but with two key differences. First, we consider repeated unknown attacks with attacker's type sampled from an unknown distribution. Second, we assume that the defender obtains limited feedback about potential attacks at the end of each period. The defender's goal is to minimize the long-term cumulative loss. Our objective is to derive an adaptive defense policy that has a low regret compared with the optimal periodic defense policy when the attack time distribution is known. A key observation is that the set of defense periods that the defender can choose from are dependent in the sense that the loss from one defense period may reveal the potential loss from other periods, especially shorter ones. Moreover, two defense policies played for the same number of rounds may span different lengths of time, which has to be taken into account when comparing the policies. In this paper, we model the defender's problem as a time associate stochastic bandit problem with dependent arms, where each arm corresponds to one possible defense period. We derive optimal defense strategies for both the finite-armed bandit setting where the defense periods can only take a finite set of values, and the continuum-armed bandit setting where the defense periods can take any values from a non-empty interval.

Our main contributions can be summarized as follows.

- We propose a stochastic time associative bandit model for optimal timing of security updates in the face of unknown attacks. Our model captures both the limited feedback about stealthy attacks and the dependence between different defense options.
- We derive upper confidence bound (UCB) based policies for time associative bandits with dependent arms. Our policies achieve a regret of $O\left(\log(T(K+1))+K\right)$ for the finite arm case, where T is the number of rounds played and K is the number of arms, and a regret of $O(T^{2/3})$ for the continuous arm setting.

Our learning model and algorithms are built upon the assumption that the defender can learn from frequent system compromises. This is reasonable for many online systems such as large online social networks and content providers and large public clouds, in which many customers are subject to similar attacks. In this setting, even if a single user is compromised occasionally, the system administrator can pool data collected from multiple users to obtain a reliable estimate quickly. For example, given the large number of attacks towards its users, Facebook can collect data from thousands of incidents of similar compromises in a short time. Our online learning algorithms can be used by Facebook to alert users to update their passwords when necessary.

Related Work

Time-related tactical security choices have been studied since the cold war era (Blackwell 1949). However, the study of timing decisions in the face of continuous and stealthy attacks is relatively new. In particular, the FlipIt game (van Dijk et al. 2013) and its variants (Laszka, Johnson, and Grossklags 2013; Laszka et al. 2014) are among the few models that study this problem in a rigorous way. However, all of these models assume that the parameters about the attacker are known to the defender at the beginning of the game. A gradient-based Bayesian learning algorithm was recently proposed in (Tan and Xia 2016) for a setting similar to ours, where the failure time was assumed to follow a Weibull distribution with one unknown parameter. In contrast, we consider a general attack time distribution.

Multi-armed bandit problems have been extensively studied for both the stochastic setting and the adversarial setting (Bubeck and Cesa-Bianchi 2012). Many variants of bandit models have been considered including bandits with side observations (Caron et al. 2012; Buccapatnam, Eryilmaz, and Shroff 2014). In the context of cybersecurity, bandit models have been applied to anomaly detection (Liu, Zhao, and Swami 2013) and stackelberg security games (Balcan et al. 2015). However, the only previous work that studies the time associative bandit model is (György et al. 2007), where the arms are assumed to be mutually independent. In contrast, we propose to model the optimal timing problem in cybersecurity as a time associative bandit problem with dependent arms and study algorithms that can exploit side-observations to improve performance.

Model

We consider the following variant of the FlipIt game (van Dijk et al. 2013) with two players, a defender and an attacker, and a security sensitive resource to protect. The at-

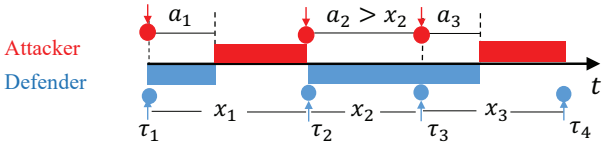


Figure 1: An example of the proposed game model. Blue circles and red circles represent the defender's and the attacker's actions, respectively. A blue segment denotes an interval where the resource is under protection, and a red segment denotes an interval where the resource is compromised.

tacker is persistent in compromising the resource. In response, the defender updates its security measures, e.g., keys, passwords, etc., from time to time to thwart the attacker. Assume a continuous time horizon. At any time instance, either the defender or the attacker can make a move to take over the resource at some cost. At time τ, the resource is under the control of the player that makes the last move before τ. Let $\tau_t, t = 1, 2, ...$ denote the time instance of the t-th defense action, and $x_t = \tau_{t+1} - \tau_t$ the t-th defense period. We assume $\tau_1 = 0$ without loss of generality. Each defense action (move) incurs a fixed cost C_D, which is known to the defender. Let X denote the set of all possible defense periods. We assume that $X \subseteq [x_{\min}, x_{\max}]$ with $x_{\min} > 0$. See Figure 1 for an example.

We further make the following assumptions about the game: (1) The attack in round t takes a random time a_t to succeed, which is $i.i.d.$ sampled from a distribution F_a that is initially unknown to the defender. In contrast, the defender recaptures the resource immediately once it makes a move, which is a reasonable assumption as it is usually much more time consuming to compromise a resource than updating its security measures. We may also interpret a_t as the awareness time that the attacker takes to discover a new vulnerability in the system. We assume that a_t is out of the control of the attacker but its distribution is known to the attacker. The attacker does not know the value of a_t until it successfully compromises the system in round t. (2) Whenever the defender makes a move, this fact is learned by the attacker immediately. On the other hand, the defender has delayed and incomplete feedback in the following sense. First, the defender only gets feedback at the end of each round. Second, at time τ_{t+1}, the defender learns the value of a_t if $a_t < x_t$, that is only when an attack is observed. Thus, the game has asymmetric feedback, a common scenario in cybersecurity. (3) The attacker is myopic and does not have a move cost. Therefore, it always attacks immediately right after a security update. Our model and solutions can be extended to the case when the attacker is myopic but with a hidden move cost (see the Appendix).

Under the above assumptions, in each round t, the resource is fully protected if $a_t \geq x_t$ and is compromised for a duration of $x_t - a_t$ otherwise. The loss to the defender in round t is then defined as:

$$l(x_t, a_t) = f[(x_t - a_t)^+] + c_d \tag{1}$$

where $f(\cdot)$ models the loss from attack and c_d models the cost of each defense action. We assume that $f(\cdot) \in [0, 1]$, $f(0) = 0$, and $f(\cdot)$ is increasing. For instance, we can consider (1) a binary loss function where $l(x_t, a_t) = 1 + c_d$ if $x_t > a_t$ and $l(x_t, a_t) = c_d$ otherwise; or (2) a linear loss function $l(x_t, a_t) = \frac{(x_t - a_t)^+}{x_{\max}} + c_d$ (where the $x_{\max}$ factor is introduced to normalize the loss value).

The defender's objective is to minimize the long-term average loss defined as follows:

$$\lambda^u = \limsup_{T\to\infty} \frac{\mathbb{E}(\sum_{t=1}^{T} l(x_t^u, a_t))}{\sum_{t=1}^{T} x_t^u} \tag{2}$$

where u denotes any defense policy and x_t^u is the t-th defense period chosen by policy u. Let $l(x) = \mathbb{E}_{a_1}(l(x, a_1))$ denote the expected loss of defense period x, and let $\lambda(x) = \frac{l(x)}{x}$ denote the time average loss of a *periodic* policy with period x. We make two observations: (1) defending all the time is not necessarily a good option as it may incur a very high defense cost; (2) it can be shown that the periodic defense policy with period $x^* = \min_{x\in X} \lambda(x)$ minimizes the long-term time average loss (Puterman 1994; György et al. 2007). However, this optimal policy cannot be found when the distribution of a_t is unknown. Let $\lambda^* = \lambda(x^*)$ denote the optimal loss. To find an optimal defense policy when the distribution of a_t is unknown, we adopt the time associative bandit model (György et al. 2007) by considering each defense period as an arm. For a defense policy $\{x_t\}$, the (pseudo) regret for the first T rounds with respect to the optimal periodic policy can be defined as:

$$\overline{R}_T = \max_{x\in X} \mathbb{E}\left[\sum_{t=1}^{T} l(x_t, a_t) - \lambda(x) x_t\right] \tag{3}$$

$$= \sum_{t=1}^{T} l(x_t) - \lambda^* \sum_{t=1}^{T} x_t \tag{4}$$

Our objective is to find a defense policy with low regret. Note that any learning algorithm with $\lim_{T\to\infty} \frac{\overline{R}_T}{T} = 0$ minimizes the long-term loss as $T \to \infty$. We also note that even if $l(x_t, a_t)$ as a function of x_t (for a fixed a_t) has a simple structure, the mean loss function $l(x_t) \triangleq E_{a_t}(l(x_t, a_t))$ may have a complicated form depending on the distribution of a_t. Therefore, previous works on linear and convex bandits cannot be directly applied to our problem. On the other hand, we observe that the defender may obtain side-observations during the game, which can be utilized to design more efficient learning algorithms.

Side observations: As we discussed before, the defender learns the value of a_t if $a_t < x_t$ (hence its loss as well) at the end of each round. From this feedback, the defender may get side observations in the following sense. Consider any round t. If $a_t < x_t$, then the defender learns the value of a_t; therefore, it learns $l(x_i, a_t)$ for any $x_i \in X$ if it has played x_i instead of x_t. On the other hand, if $a_t \geq x_t$, the defender only learns the value of $l(x_i, a_t) = c_d$ for any $x_i \leq x_t$, but not the value of $l(x_j, a_t)$ for $x_j > x_t$. This implies that playing an arm that corresponds to a longer defense period provides more side observations about other arms. Our learning algorithm incorporates these side-observations to minimize the

expected regret. Indeed, our algorithm and its regret bound apply to any loss function $l(x_t, a_t)$ where playing one period provides side-observations to all shorter periods.

Multiple resources: Our model can be readily extended to consider multiple resources (nodes) subject to $i.i.d.$ attacks, which can be used to model multiple users subject to independent attacks in an online system such as Facebook. In this case, samples from multiple nodes can be pooled together when choosing the next defense period for a node. Consider a system with N nodes that are subject to $i.i.d.$ attacks with unknown attack times sampled from F_a. Let τ_{st} denote the time instance of the t-th security update on node s and $x_{st} = \tau_{s(t+1)} - \tau_{st}$ the t-th defense period for node s. Note that x_{st} can be different for different s. Let a_{st} denote the attack time in the t-th attack towards node s. Let $l(x_{st}, a_{st})$ denote the loss to the defender in round t over node s. When the nodes are subject to $i.i.d.$ attacks, there is an optimal defense period x^* for all nodes with minimum time average loss λ^*, similar to the single node setting. The $i.i.d.$ assumption may hold in practice because (1) some parameters such as the attack time may be out of the control of the attacker and can be approximated as $i.i.d.$ random variables during the time horizon of the game; (2) an adversarial attacker may choose to avoid correlated attacks to make its behavior more unpredictable. Assume that the game is played for T_s rounds over node s, and let $T = \sum_s T_s$. Then the regret over the T rounds of play across all the nodes can be defined as $\overline{R}_T = \sum_{s=1}^{N} \sum_{t=1}^{T_s} l(x_{st}) - \lambda^* \sum_{s=1}^{N} \sum_{t=1}^{T_s} x_{st}$. Note that, when choosing x_{st}, feedback from all the nodes received before τ_{st} can be used. Our online learning algorithms can be directly applied to this setting.

Optimal Timing Algorithms

In this section, we present our learning algorithms for the optimal timing problem for both discrete and continuous defense periods.

Discrete Defense Periods

We first consider the finite-armed setting where the set of defense periods is finite, denoted by $X = \{x_1, ..., x_K\}$. Let i_t denote the index of the arm played in round t, i.e., $x_{i(t)}$ is the defense period chosen for round t. Let $n_i(t) = \sum_{s=1}^{t} \mathbb{I}(i_s = i)$ denote the number of plays of arm i during the first t rounds. Let $\bar{l}_{i,t} = \frac{1}{n_i(t)} \sum_{s=1}^{t} \mathbb{I}(i_s = i) l(x_{i(s)}, a_s)$ denote the average loss from arm i during the first t rounds, and $\overline{\lambda}_{i,t} = \frac{\bar{l}_{i,t}}{x_i}$ the time average loss of arm i. To simplify the notation, we omit the subscript t in $\bar{l}_{i,t}$ and $\overline{\lambda}_{i,t}$ when it is clear from the context, and let $l_i \triangleq l(x_i)$ and $\lambda_i \triangleq \lambda(x_i)$. Let $\Delta_i = l_i - x_i \lambda^*$ denote the *relative* loss of playing arm i. Note that $\Delta_{i^*} = 0$ for an optimal arm i^* and $\Delta_i \in [0, 1]$ by our assumption about l. Then $\overline{R}_T = \sum_{i \neq i^*} \Delta_i \mathbb{E}(n_i(T))$. Let $\Delta_{\min} \triangleq \min_{i:\Delta_i > 0} \Delta_i$ and $\Delta_{\max} \triangleq \max_i \Delta_i$.

To derive an optimal defense policy, we consider the following variant of the improved upper confidence bound based policy proposed in (Auer and Ortner 2010) for stochastic bandits. We modify the improved UCB policy to address the time associative regret while taking the dependence between arms into account.

Algorithm 1 Improved UCB algorithm for time-associative bandits with side observations

Input: A set of periods X, the number of rounds T.
Initialization: Set $\tilde{\Delta}_0 = 1, X_0 = X$.
for $m = 0, 1, 2, ...,$ **do**
$x_{(1)} = \min\{x_i \in X_m\}$; $x_{(2)} = \max\{x_i \in X_m\}$.
Arm selection:
If $|X_m| = 1$, play the single period in X_m until T.
Else play the longest period in X_m until round $\min(n_m, T)$, where $n_m = \left\lceil \frac{2\gamma_m \log(T(K+1)\tilde{\Delta}_m^2)}{\tilde{\Delta}_m^2} \right\rceil$ and $\gamma_m = \left(1 + \frac{x_{(2)}}{x_{(1)}}\right)^2$; update $\bar{l}_i, \overline{\lambda}_i$ for all $x_i \in X_m$.
Arm elimination:
$\overline{\lambda}_m = \min_{x_i \in X_m} \left(\overline{\lambda}_i + c_m / x_i\right)$ where $c_m = \sqrt{\frac{\log(T(K+1)\tilde{\Delta}_m^2)}{2n_m}}$.
To get X_{m+1}, delate all the periods $x_i \in X_m$ such that $\bar{l}_i - x_i \overline{\lambda}_m \geq \min_{x_j \in X_m} \bar{l}_j - x_j \overline{\lambda}_m + 2\left(1 + \frac{x_j}{x_{(1)}}\right) c_m$.
Reset: $\tilde{\Delta}_{m+1} = \frac{\tilde{\Delta}_m}{2}$.

The algorithm proceeds in multiple stages, where each stage involves multiple rounds (see Algorithm 1). In each stage m, as in the improved UCB policy, our policy estimates Δ_i by a value $\tilde{\Delta}_m$, and maintains a set of active arms X_m. $\tilde{\Delta}_0$ is initialized to 1 and is halved in each stage. X_0 initially contains all the arms. At the end of each stage m, a subset of arms are deleted from X_m according to their observed losses in previous rounds. Compared with the improved UCB policy for stochastic bandits, our policy has several key differences. First, in the arm selection phase, each active arm is played $n_m - n_{m-1}$ times in stage m in the improved UCB policy, where n_m is a function of $\tilde{\Delta}_m$ and is chosen so that any suboptimal arm i is eliminated as soon as $\tilde{\Delta}_m < \frac{\Delta_i}{2}$ with high probability. In contrast, only the longest period in X_m is played $n_m - n_{m-1}$ times in our policy, which provides side observations to all the shorter periods as we discussed above. For any arm $x_i \in X_m$, $\bar{l}_i$ is defined as if i is played in all the previous n_m rounds. In addition, the definition of n_m in our policy is different from the improved UCB policy. In particular, n_m depends on the ratio of the maximum active period to the minimum active period, which is needed to bound the time associative regret. Second, in the arm elimination phase, we compare the relative losses of arms instead of average losses as in the improved UCB, since average loss alone does not take the length of a defense period into account. In particular, we estimate the relative loss of arm i by $\bar{l}_{i,n_m} - x_i \overline{\lambda}_m$, where $\overline{\lambda}_m$ is an estimate of λ^* defined by

$$\overline{\lambda}_m = \min_{x_i \in X_m} \left(\overline{\lambda}_{i,n_m} + c_m / x_i\right) \tag{5}$$

where $c_m = \sqrt{\frac{\log(T(K+1)\tilde{\Delta}_m^2)}{2n_m}}$. The value of c_m is chosen

so that for all i, $\overline{\lambda}_{i,n_m}$ is in the c_m/x_i-vicinity of λ_i with high probability.

To establish the regret bound of the algorithm, we need the following lemmas.

Lemma 1. *(Chernoff-Hoeffding Bound (Hoeffding 1963)) Let $X_1, X_2, ..., X_n$ be a sequence of independent random variables with support $[a, b]$ and $E(X_t) = \mu$ for all X_t. Let $\overline{X}_n = \frac{1}{n}\sum_{t=1}^{n} X_t$. Then for any $\epsilon > 0$, we have*

$$\mathbb{P}\{\overline{X}_n \geq \mu + \epsilon\} \leq e^{-\frac{2n\epsilon^2}{(b-a)^2}},$$
$$\mathbb{P}\{\overline{X}_n \leq \mu - \epsilon\} \leq e^{-\frac{2n\epsilon^2}{(b-a)^2}}.$$

Lemma 2. *Consider any stage m where there is an optimal arm $i^* \in X_m$. If $l_i \leq \bar{l}_i + c_m$ for all $x_i \in X_m$ and $l^* \geq \bar{l}_{i^*} - c_m$, then we must have $\lambda^* \leq \overline{\lambda}_m \leq \lambda^* + 2c_m/x_{i^*}$.*

Proof. To see this, let $x_j \in X_m$ be the arm that minimizes $\overline{\lambda}_j$. Then we have $\lambda^* \leq \lambda_j \leq \overline{\lambda}_j + c_m/x_j = \overline{\lambda}_m$, and $\overline{\lambda}_m \leq \overline{\lambda}_{i^*} + c_m/x_{i^*} \leq \lambda^* + 2c_m/x_{i^*}$. □

We now show the following bound on the expected regret of Algorithm 1.

Theorem 1. *The expected regret of Algorithm 1 is at most $\frac{48\gamma\log\left(T(K+1)\frac{\Delta_{\max}^2}{4}\right)}{\Delta_{\min}} + \sum_{i:\Delta_i>0}\left(\Delta_i + \frac{48}{\Delta_i}\right)$, where $\gamma = (1 + \frac{x_{\max}}{x_{\min}})^2$.*

Proof. Without loss of generality, we assume that the optimal arm is unique and has index K, and sort the set of arms such that $\Delta_1 \geq \Delta_2 \geq \cdots \geq \Delta_K = 0$. For any suboptimal arm i, let $m_i = \min\{m : \tilde{\Delta}_m < \frac{1}{2}\Delta_i\}$ denote the first stage in which $\tilde{\Delta}_m < \frac{1}{2}\Delta_i$. We have $2^{m_i} = \frac{1}{\tilde{\Delta}_{m_i}} \leq \frac{4}{\Delta_i} < \frac{1}{\tilde{\Delta}_{m_i+1}} = 2^{m_i+1}$. Note that $m_1 \leq m_2 \leq \cdots \leq m_{K-1}$.

We consider the following events similar to (Perchet and Rigollet 2013). Let A_i denote the event that the optimal arm has *not* been eliminated before stage m_i, and B_i the event that every arm $j \in \{1, 2, ..., i\}$ has been eliminated in stage m_j or before. We have $A_1 \supseteq A_2 \cdots \supseteq A_{K-1}$ and $B_1 \supseteq B_2 \cdots \supseteq B_{K-1}$. Let $C_i = A_i \cap B_i$ for $i \in \{1, 2, ..., K-1\}$. Under the event C_i, let U_i denote the contribution to the regret from arms $\{1, 2, ..., i\}$ and V_i the contribution to the regret from arms $\{i+1, ..., K-1\}$. We observe that $V_i \leq T\Delta_{i+1}$. Let C_0 denote the sample space. We then have

$$\begin{aligned}
\overline{R}_T &= \sum_{i=1}^{K-1} \mathbb{P}(C_{i-1}\backslash C_i)(U_i + V_i) \\
&\leq \sum_{i=1}^{K-1} U_i\mathbb{P}(C_{i-1}\backslash C_i) + \sum_{i=1}^{K-1} T\Delta_i\mathbb{P}(C_{i-1}\backslash C_i) \\
&\leq \sum_{i=1}^{K-1} U_i\mathbb{P}(C_{i-1}\backslash C_i) + \sum_{i=1}^{K-1} T\Delta_i\mathbb{P}(B_i^c \cap B_{i-1} \cap A_i) \\
&\quad + \sum_{i=1}^{K-1} T\Delta_i\mathbb{P}(A_i^c \cap C_{i-1}) \qquad (6)
\end{aligned}$$

We bound each of the three terms in (6) as follows:

First term in (6): Under the event C_i, each suboptimal arm $j \in \{1, 2, ..., i\}$ is eliminated on or before round $n_{m_j} = \left\lceil \frac{2\gamma_{m_j}\log(T(K+1)\tilde{\Delta}_{m_j}^2)}{\tilde{\Delta}_{m_j}^2} \right\rceil$. Among these arms, let $j_1, j_2, ..., j_k$ denote the sequence of suboptimal arms played where $x_{j_1} > ... > x_{j_k}$, and arm j_i is eliminated in stage $m'_{j_i} \leq m_{j_i}$. Let $B \triangleq 2\gamma\log\left(T(K+1)\frac{\Delta_{\max}^2}{4}\right)$. We then have

$$\begin{aligned}
U_i &\leq \Delta_{j_1} n_{m'_{j_1}} + \sum_{i=2}^{k} \Delta_{j_i}(n_{m'_{j_i}} - n_{m'_{j_{i-1}}}) \\
&\leq \Delta_{j_1} n_{m'_{j_1}} + \sum_{i=2}^{k} \Delta_{j_i}\left(1 + \frac{2\gamma\log(T(K+1)\tilde{\Delta}_{m_{j_i}}^2)}{\tilde{\Delta}_{m_{j_i}}^2} \right. \\
&\qquad \left. - \frac{2\gamma\log(T(K+1)\tilde{\Delta}_{m_{j_{i-1}}}^2)}{\tilde{\Delta}_{m_{j_{i-1}}}^2}\right) \\
&\leq \Delta_{j_1} n_{m'_{j_1}} + \sum_{i=2}^{k} \Delta_{j_i} + 4B\sum_{i=2}^{k} \tilde{\Delta}_{m_{j_i}}\left(\frac{1}{\tilde{\Delta}_{m_{j_i}}^2} - \frac{1}{\tilde{\Delta}_{m_{j_{i-1}}}^2}\right) \\
&\leq \Delta_{j_1} n_{m'_{j_1}} + \sum_{i=2}^{k} \Delta_{j_i} + 4B\sum_{i=2}^{k} \frac{1.5\tilde{\Delta}_{m_{j_{i-1}}}(\tilde{\Delta}_{m_{j_{i-1}}} - \tilde{\Delta}_{m_{j_i}})}{\tilde{\Delta}_{m_{j_i}}\tilde{\Delta}_{m_{j_{i-1}}}^2} \\
&= \Delta_{j_1} n_{m'_{j_1}} + \sum_{i=2}^{k} \Delta_{j_i} + 6B\sum_{i=2}^{k}\left(\frac{1}{\tilde{\Delta}_{m_{j_i}}} - \frac{1}{\tilde{\Delta}_{m_{j_{i-1}}}}\right) \\
&\leq \sum_{i=1}^{k} \Delta_{j_i} + 6B\frac{1}{\tilde{\Delta}_{m_{j_k}}} \leq \sum_{i=1}^{k} \Delta_{j_i} + 24B\frac{1}{\Delta_{j_k}}
\end{aligned}$$

Therefore, $\sum_{i=1}^{K-1} U_i\mathbb{P}(C_{i-1}\backslash C_i) \leq \sum_{i=1}^{K-1}\Delta_i + 24B\frac{1}{\Delta_{\min}}$.

Second term in (6): Under the event $B_i^c \cap B_{i-1} \cap A_i$, the optimal arm is not eliminated by m_i, neither does arm i. We first note that if $\bar{l}_i \geq l_i - c_{m_i}$ and $\bar{l}_K \leq l_K + c_{m_i}$ hold, then arm i will be eliminated in round m_i. Indeed, from the definitions of c_m and n_m, we have $c_{m_i} \leq \frac{\tilde{\Delta}_{m_i}}{2\sqrt{\gamma_{m_i}}} = \frac{\tilde{\Delta}_{m_{i+1}}}{\sqrt{\gamma_{m_i}}} < \frac{\Delta_i}{4\sqrt{\gamma_{m_i}}}$. Then from Lemma 2, we have

$$\begin{aligned}
\bar{l}_i - x_i\overline{\lambda}_{m_i} &\geq \bar{l}_i - x_i(\lambda^* + 2c_{m_i}/x_K) \\
&\geq l_i - x_i(\lambda^* + 2c_{m_i}/x_K) - c_{m_i} \\
&= l_K - x_K\lambda^* + \Delta_i - 2\frac{x_i}{x_K}c_{m_i} - c_{m_i} \\
&> l_K - x_K\lambda^* + 4\sqrt{\gamma_{m_i}}c_{m_i} - 2\frac{x_i}{x_K}c_{m_i} - c_{m_i} \\
&\geq \bar{l}_K - x_K\lambda^* + 2\left(1 + \frac{x_K}{x_{(1)}}\right)c_{m_i} \\
&\geq \bar{l}_K - x_K\overline{\lambda}_{m_i} + 2\left(1 + \frac{x_K}{x_{(1)}}\right)c_{m_i}
\end{aligned}$$

where $x_{(1)}$ is the minimum active period in stage m_i. It follows that arm i is eliminated in stage m_i as claimed.

It follows that $\mathbb{P}(B_i^c \cap B_{i-1} \cap A_i) \leq \mathbb{P}(\bar{l}_i < l_i - c_{m_i}) + \mathbb{P}(\bar{l}_{i^*} > l_{i^*} + c_{m_i}) \leq \frac{1}{T(K+1)\tilde{\Delta}_{m_i}^2} + \frac{1}{T(K+1)\tilde{\Delta}_{m_i}^2} \leq \frac{1}{T\tilde{\Delta}_{m_i}^2}$ by the Chernoff-Hoeffding bound. Therefore, the second term in (6) can be bounded by $\sum_i T\Delta_i\frac{1}{T\tilde{\Delta}_{m_i}^2} \leq \sum_i \frac{16}{\Delta_i}$.

Third term in (6): Under the event $A_i^c \cap C_{i-1}$, every arm $j \in \{1, 2, ...i-1\}$ has been eliminated by stage m_j and the optimal arm

is eliminated by some arm $k \geq i$ in some stage m_* where $m_{i-1} < m_* \leq m_i$. We first claim that if $\bar{l}_k \geq l_k - c_{m_*}$ and $\bar{l}_K \leq l_K + c_{m_*}$ hold, then the optimal arm is *not* eliminated by arm k in stage m_*. To see this, assume that the optimal arm is eliminated, which happens only when $\bar{l}_K - x_K\bar{\lambda}_{m_*} \geq \bar{l}_k - x_k\bar{\lambda}_{m_*} + 2\left(1 + \frac{x_k}{x_{(1)}}\right)c_{m_*}$. From Lemma 2, we have:

$$\begin{aligned}
l_k - x_k\lambda^* \leq & \bar{l}_k + c_{m_*} - x_k(\bar{\lambda}_{m_*} - 2c_{m_*}/x_K) \\
\leq & \bar{l}_k - x_k\bar{\lambda}_{m_*} + c_{m_*} + 2\frac{x_k}{x_K}c_{m_*} \\
\leq & \bar{l}_k - x_k\bar{\lambda}_{m_*} + 2\left(1 + \frac{x_k}{x_{(1)}}\right)c_{m_*} - c_{m_*} \\
\leq & \bar{l}_K - x_K\bar{\lambda}_{m_*} - c_{m_*} \\
\leq & l_K - x_K\lambda^*
\end{aligned}$$

which contradicts the fact that k is suboptimal. It follows that the probability that the optimal arm is eliminated by a fixed arm $k \geq i$ in a fixed stage $m_* \leq m_i$ is bounded by $\mathbb{P}(\bar{l}_k < l_k - c_{m_*}) + \mathbb{P}(\bar{l}_K > l_K + c_{m_*}) \leq \frac{1}{T\tilde{\Delta}^2_{m_*}}$ by the Chernoff-Hoeffding bound. Therefore, the third term in (6) is bounded by

$$\begin{aligned}
& \sum_{i=1}^{K-1}\sum_{m_*=m_{i-1}+1}^{m_i}\sum_{k=i}^{K-1}\frac{1}{T\tilde{\Delta}^2_{m_*}}T\Delta_i \\
= & \sum_{m_*=0}^{\max_i m_i}\sum_{k:m_k\geq m_*}\frac{1}{T\tilde{\Delta}^2_{m_*}}T\max_{h:m_h\geq m_*}\Delta_h \\
\leq & \sum_{m_*=0}^{\max_i m_i}\sum_{k:m_k\geq m_*}\frac{1}{\tilde{\Delta}^2_{m_*}}4\tilde{\Delta}_{m_*} \\
= & \sum_{i=1}^{K-1}\sum_{m_*=0}^{m_i}\frac{4}{\tilde{\Delta}_{m_*}} \leq \sum_{i=1}^{K-1}4\cdot 2^{m_i+1} \leq \sum_{i=1}^{K-1}\frac{32}{\Delta_i}
\end{aligned}$$

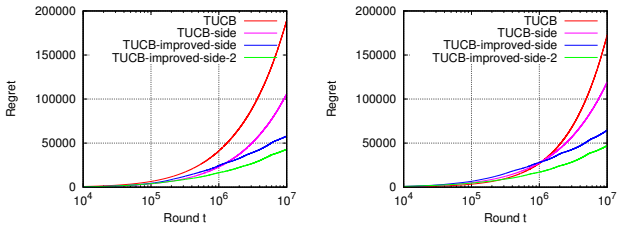
Putting all the three cases together, we get the desired regret bound. □

Remark 1. *Our algorithm achieves a regret where the coefficient of the* $\log(T)$ *term is independent of* K*, the number of arms. This is obtained by utilizing the side observations among arms. In contrast, a direct application of the UCB based policy for time-associative bandits in (György et al. 2007) to our problem leads to a regret of* $O\left(\sum_{i=1}^{K}\frac{\gamma\log(T(K+1))}{\Delta_i^2}\right)$*, where the* $\log(T)$ *term has a coefficient that is linear of* K*.*

Remark 2. *In our numerical study, we also consider a variant of Algorithm 1 where in arm elimination phase, we delate all the periods* $x_i \in X_m$ *such that* $\bar{l}_i - x_i\bar{\lambda}_m \geq \min_{x_j\in X_m}\bar{l}_j - x_j\bar{\lambda}_m + 4c_m$*. By using a smaller confidence interval, this variant eliminates suboptimal arms more aggressively than Algorithm 1. Although we are not able to prove a regret bound for this variant, it exhibits even better performance than Algorithm 1 in our numerical study.*

Continuous Defense Periods

We next consider the case where the defense periods can take any real value in $X = [x_{\min}, x_{\max}]$. Note that in this case, the bound given in Theorem 1 can be very poor due to the large K and small Δ_i. Built upon Algorithm 1, we propose a new policy with a regret that is independent of K and Δ_i under the following assumption. Let $l(x) = E_{a_1}(l(x, a_1))$ denote the expected loss when a period x is played. We assume that $l(x)$ is Lipschitz continuous: there exists a constant $L \geq 0$ such that for any $x_1, x_2 \in X$, $|l(x_1) - l(x_2)| \leq L|x_1 - x_2|$. For instance, when the attack time follows a uniform distribution in $[a_1, a_2]$, and $f(\cdot)$ is binary, we have $|l(x_1) - l(x_2)| \leq \frac{1}{a_2-a_1}|x_1 - x_2|$, and we can take $L = \frac{1}{a_2-a_1}$.



(a) binary loss (b) linear loss

Figure 2: Numerical Results.

Our algorithm is inspired by UCB for continuous bandits (UCBC) (Auer, Ortner, and Szepesvári 2007). We first divide X into n subintervals of equal length, where n is a parameter to be determined (see Algorithm 2). Let $x_k \triangleq x_{\min} + k\frac{x_{\max}-x_{\min}}{n}$ denote the longest period in the k-th interval. We then apply Algorithm 1 to the set of arms $I \triangleq \{x_1, ..., x_n\}$.

Algorithm 2 Improved UCB based optimal timing with continuous periods

Input: A set of periods $X = [x_{\min}, x_{\max}]$, the number of rounds T, the number of subintervals n.
Initialization: For $k = 1, 2, ..., n$, $x_k = x_{\min} + k\frac{x_{\max}-x_{\min}}{n}$.
Apply Algorithm 1 to $I = \{x_1, x_2..., x_n\}$.

Define $I_1 \triangleq [x_{\min}, x_1]$ and $I_k \triangleq (x_{k-1}, x_k]$ for $1 < k \leq n$. For any $x \in X$, let $\Delta(x) \triangleq l(x) - x\lambda^*$ denote the relative loss of x. Among $x_k \in I$, assume x_{k^*} has the minimum $\lambda(x_k)$. Let $\Delta'(x) \triangleq l(x) - x\lambda(x_{k^*})$ denote the relative loss of arm x with respect to x_{k^*}. It is clear that $\Delta'(x) \leq \Delta(x)$. We further have the following property about Δ and Δ'.

Lemma 3. $\Delta(x_{k^*}) \leq L'n^{-1}$ *and* $\Delta(x) - \Delta'(x) \leq L'n^{-1}$*, where* $L' = L\frac{x_{\max}(x_{\max}-x_{\min})}{x_{\min}}$.

From the lemma, we can establish the following performance bound for Algorithm 2.

Theorem 2. *The expected regret of the variant of the improved UCB policy for continuous bandits described in Algorithm 2 is at most* $3L'n^{-1}T + \frac{48\gamma\log(T(n+1))}{L'n^{-1}} + \frac{48n^2}{L'} + n\Delta_{\max}$*. By taking* $n = T^{1/3}$*, we have* $R_T \leq O(T^{2/3})$.

Proofs of Lemma 3 and Theorem 2 are provided in the Appendix.

Numerical Results

In this section, we demonstrate the advantages of our learning algorithms through numerical study. We use the following synthetic dataset. We assume that the attack time a_t follows an $i.i.d.$ Weibull Distribution with CDF $F(a) = 1 - e^{-(a/\lambda)^b}$ for $a \geq 0$ and $F(a) = 0$ for $a < 0$. This model has been used in reliability engineering (Mazzuchi and Soyer 1996) and cybersecurity (Tan and Xia 2016) to model failure times. Note that when $b = 1$, the Weibull Distribution becomes the exponential distribution. By setting $b > 1$, the model indicates that the failure rate increases with time. We set $b = 2$ in experiments. In each trial, λ is chosen from the interval $[1, 20]$ uniformly at random. We consider a 19 arm setting with x_i evenly distributed in $[1, 10]$ with a step size of 0.5. We consider both the binary loss function and the linear loss function mentioned in the model section. In both cases, we fix the defense cost to $c_d = 0.1$. With these parameter settings, we observe that the best arm varies over the feasible defense periods when we vary λ.

We focus on the case where side observations are available (without attack cost) and compare our algorithms with the UCB based time-associative bandit algorithm in (György et al. 2007) (TUCB) that do not consider side observations. We further consider a variant of TUCB that uses the TUCB policy to choose the arm to play in each round and obtains side-observations after each play (TUCB-side). This algorithm can be considered as the application of the UCB-N policy and the UCB-MaxN policy in (Caron et al. 2012) to the time associative bandit model (UCB-N and UCB-MaxN give the same policy under the dependence structure we consider.) For our algorithms, we evaluate both Algorithm 1 (TUCB-improved-side) and its variant discussed above (TUCB-improved-side-2). The results are averaged over 100 independent trials and are given in Figure 2. We note that the linear loss setting represents the harder case since it introduces smaller variances across arms. We observe that for both loss functions, our algorithms can significantly reduce long-term regrets compared to TUCB and TUCB-side by carefully incorporating side observations. Moreover, TUCB-improved-side-2 achieves the best performance among the four algorithms.

Acknowledgments

This research was supported in part by a grant from the Army Research Office AROW911NF-15-1-0277, and an Army Research Office MURI W911NF-12-1-0385.

References

Auer, P., and Ortner, R. 2010. UCB revisited: Improved regret bounds for the stochastic multi-armed bandit problem. *Periodica Mathematica Hungarica* 61(1):55–65.

Auer, P.; Ortner, R.; and Szepesvári, C. 2007. Improved Rates for the Stochastic Continuum-Armed Bandit Problem. In *Proc. of COLT*.

Balcan, N.; Blum, A.; Haghtalab, N.; and Procaccia, A. 2015. Commitment Without Regrets: Online Learning in Stackelberg Security Games. In *Proc. of EC*.

Barnett, E. 2011. Hackers go after Facebook sites 600,000 times every day. *The Telegraph, Oct.29, 2011*.

Beattie, S.; Arnold, S.; Cowan, C.; Wagle, P.; Wright, C.; and Shostack, A. 2002. Timing the Application of Security Patches for Optimal Uptime. In *Proc. of USENIX LISA*.

Blackwell, D. 1949. The noisy duel, one bullet each, arbitrary accuracy. Technical report, The RAND Corporation, D-442.

Bowers, K. D.; Dijk, M. E. V.; Juels, A.; Oprea, A. M.; Rivest, R. L.; and Triandopoulos, N. 2014. Graph-based approach to deterring persistent security threats. US Patent 8813234.

Bubeck, S., and Cesa-Bianchi, N. 2012. Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems. *Foundations and Trends in Machine Learning* 5(1):1–122.

Buccapatnam, S.; Eryilmaz, A.; and Shroff, N. B. 2014. Stochastic bandits with side observations on networks. In *Proc. of Sigmetrics*.

Caron, S.; Kveton, B.; Lelarge, M.; and Bhagat, S. 2012. Leveraging Side Observations in Stochastic Bandits. In *Proc. of UAI*.

György, A.; Kocsis, L.; Szabó, I.; and Szepesvári, C. 2007. Continuous Time Associative Bandit Problems. In *Proc. of IJCAI*.

Hoeffding, W. 1963. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association* 58(301):13–30.

Juels, A.; Dijk, M. E. V.; Oprea, A. M.; and Rivest, R. L. 2016. Scheduling of defensive security actions in information processing systems. US Patent 9471777.

Laszka, A.; Horvath, G.; Felegyhazi, M.; and Buttyán, L. 2014. Flipthem: Modeling targeted attacks with flipit for multiple resources. In *Proc. of GameSec*.

Laszka, A.; Johnson, B.; and Grossklags, J. 2013. Mitigating Covert Compromises: A Game-Theoretic Model of Targeted and Non-Targeted Covert Attacks. In *Proc. of WINE*.

Liu, K.; Zhao, Q.; and Swami, A. 2013. Dynamic Probing for Intrusion Detection under Resource Constraints. In *Proc. of ICC*.

Mazzuchi, T. A., and Soyer, R. 1996. A Bayesian perspective on some replacement strategies. *Reliability Engineering and System Safety* 51(3):295–303.

Perchet, V., and Rigollet, P. 2013. The multi-armed bandit problem with covariates. *The Annals of Statistics* 41(2):693–721.

Puterman, M. L. 1994. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley-Interscience.

Tan, Y., and Xia, C. H. 2016. Cyber Maintenance Policy Optimization via Adaptive Learning. In *Proc. of Infocom*.

van Dijk, M.; Juels, A.; Oprea, A.; and Rivest, R. L. 2013. FlipIt: The Game of "Stealthy Takeover". *Journal of Cryptology* 26(4):655–713.

Zheng, Z.; Shroff, N. B.; and Mohapatra, P. 2016. When to Reset Your Keys: Optimal Timing of Secu-

rity Updates via Learning. Technical Report, available online at www.cs.tulane.edu/~zzheng3/publication/optimal-timing-aaai17.pdf.

Appendix

Proof of Lemma 3

Assume $\lambda^* = \lambda(x^*)$ for $x^* \in I_j$. By the definition of k^*, we have $\lambda(x_{k^*}) = \frac{l(x_{k^*})}{x_{k^*}} \leq \frac{l(x_j)}{x_j} = \lambda(x_j)$. It follows that

$$\begin{aligned}\Delta(x_{k^*}) &= l(x_{k^*}) - x_{k^*}\lambda^* \\ &\leq l(x_j)\frac{x_{k^*}}{x_j} - x_{k^*}\lambda^* \\ &= \frac{x_{k^*}}{x_j}(l(x_j) - x_j\lambda^*) \\ &\leq \frac{x_{\max}}{x_{\min}}\Delta(x_j).\end{aligned}$$

Moreover,

$$\begin{aligned}\Delta(x_j) &= l(x_j) - x_j\lambda^* \\ &= l(x_j) - x_j\frac{l(x^*)}{x^*} \\ &\leq l(x_j) - l(x^*) \\ &\leq L|x_j - x^*| \\ &\leq L\frac{x_{\max} - x_{\min}}{n}\end{aligned}$$

Therefore, $\Delta(x_{k^*}) \leq L\frac{x_{\max}}{x_{\min}}\frac{x_{\max}-x_{\min}}{n}$. Similarly, $\Delta(x) - \Delta'(x) = x(\lambda(x_{k^*}) - \lambda(x^*)) \leq x(\lambda(x_j) - \lambda(x^*)) = x(\frac{l(x_j)}{x_j} - \frac{l(x^*)}{x^*}) \leq \frac{x}{x^*}(l(x_j) - l(x^*)) \leq \frac{x_{\max}}{x_{\min}}L\frac{x_{\max}-x_{\min}}{n}$.

Proof of Theorem 2

We split the set of intervals into two parts. Define $S \triangleq \{k \in I : \Delta(x_k) > 2L'n^{-1}\}$. We have $\overline{R}_T = \sum_{k=1}^{n}\Delta(x_k)E(n_k(T))$ where $n_k(T)$ denotes the number of times that x_k is played. Let $\overline{R}_T = \overline{R}_{T,1} + \overline{R}_{T,2}$, where $\overline{R}_{T,1} = \sum_{k\notin S}\Delta(x_k)E(n_k(T))$ and $\overline{R}_{T,2} = \sum_{k\in S}\Delta(x_k)E(n_k(T))$. It is easy to see that $\overline{R}_1 \leq 2L'n^{-1}T$. On the other hand, we have

$$\begin{aligned}\overline{R}_{T,2} &= \sum_{k\in S}\Delta(x_k)E(n_k(T)) \\ &\leq \sum_{k\in S}(\Delta'(x_k) + L'n^{-1})E(n_k(T)) \\ &\leq \sum_{k\in S}\Delta'(x_k)E(n_k(T)) + L'n^{-1}T\end{aligned}$$

By Theorem 1, we have $\sum_{k\in S}\Delta'(x_k)E(n_k(T)) \leq \frac{48\gamma\log(T(n+1))}{\min_{k\in S}\Delta'(x_k)} + \sum_{k\in S}\left(\Delta'(x_k) + \frac{48}{\Delta'(x_k)}\right)$. By Lemma 3, $\Delta'(x_k) \geq \Delta(x_k) - L'n^{-1} \geq L'n^{-1}$ for any $k \in S$. It follows that, $\overline{R}_T \leq 3L'n^{-1}T + \frac{48\gamma\log(T(n+1))}{L'n^{-1}} + \frac{48n^2}{L'} + n\Delta_{\max}$. By taking $n = T^{1/3}$, we have $\overline{R}_T \leq O(T^{2/3})$.

Costly Attacks

Our model and solutions can be extended to a myopic attacker with a hidden attack cost c_a. After observing the defender's move in the beginning of round t, the attacker attacks immediately if $\mathbb{E}_{a_t}[(x_t - a_t)^+] > c_a$, and does not attack in that round otherwise That is, it attacks only if its expected benefit is larger than the attack cost. Equivalently, we can assume there is a period x_0 such that the attacker attacks only when $x_t > x_0$. We can further distinguish the following two cases.

- **Fixed Attack Cost:** In this case, there is a fixed x_0 that is unknown to the defender such that there is no attack in round t if $x_t \leq x_0$ is played (thus the defender only suffers from the defense cost). The defender's loss in round t is defined as:
$$l(x_t, a_t) = \begin{cases} c_d & \text{if } x_t \leq x_0, \\ f[(x_t - a_t)^+] + c_d & \text{if } x_t > x_0. \end{cases} \quad (7)$$
Note that by setting $x_0 \leq x_{\min}$, this case reduces to the setting when there is no attack cost.
- **Random Attack Cost:** In this case, x_0 is $i.i.d.$ sampled in each round from an unknown distribution. The defender's loss in round t is same as (7).

When there is a fixed attack cost, playing a longer period x_i provides side-observation to a shorter arm x_j only when $x_j > x_0$. Since x_0 is unknown, it is insufficient to only play the longest period in each round as we did before. Therefore, we modify Algorithm 1 by maintaining a set of periods Y_m that the defender knows to be longer than x_0 by stage m. We set $Y_0 = \emptyset$. Since x_0 is fixed, whenever an attack is observed when playing a period x_i, we know $x_j > x_0$ for all $x_j \geq x_i$. In each stage m, for each active period in Y_m, the algorithm plays the longest one only as before to exploit side-observations. However, each active period not in Y_m requires further exploration. Therefore, they are also played the same number of times in each stage. But whenever an attack is observed on x_i, all $x_j \geq x_i$ are added to Y_m and they don't need to be played separately any further. We can prove the following regret bound for the modified algorithm.

Theorem 3. *The expected regret of the above algorithm is at most* $\sum_{i:x_i\leq x_0}\frac{B_i}{\Delta_i} + \sum_{i:x_i>x_0}\min\left(\frac{B_i}{\Delta_i}, \frac{\Delta_i}{p_i}\right) + \frac{48\gamma\log\left(T(K+1)\frac{\Delta_{\max}^2}{4}\right)}{\Delta_{\min}} + \sum_{i:\Delta_i>0}\left(\Delta_i + \frac{48}{\Delta_i}\right)$, *where* $B_i = 32\gamma\log\left(T(K+1)\frac{\Delta_i^2}{4}\right)$, *and* $p_i \triangleq \mathbb{P}(l(x_i, a_t) > 0)$.

Proof. We adopt a similar argument as in the proof of Theorem 1. In particular, the last two terms in (6) can be bounded using the same argument. The only difference is in the first term. For every period $x_i < x_0$, we bound its regret up to stage m_i by $\Delta_i n_{m_i} \leq \Delta_i\left\lceil\frac{2\gamma\log(T(K+1)\tilde{\Delta}_{m_i}^2)}{\tilde{\Delta}_{m_i}^2}\right\rceil < \Delta_i\left(1 + \frac{32\gamma\log(T(K+1)\frac{\Delta_i^2}{4})}{\Delta_i^2}\right)$, which gives the first term (and part of the last term) in the regret. Next consider a period $x_j > x_0$. Let $p_j \triangleq \mathbb{P}(l(x_j, a_t) > 0)$ denote the probability that an attack is observed when playing x_j. From the

algorithm, the expected number of rounds until x_j is added to Y_m is bounded by $\min(\frac{1}{p_j}, n_{m_j})$, which gives the second term in the regret (and part of the last term). After x_j is added to Y_m, it is played only when it becomes the longest active period in Y, which gives the third term in the regret using the same argument for the first term of (6) in Theorem 1. □

When x_0 is $i.i.d.$ sampled from an unknown distribution, playing longer periods do not provide deterministic side-observations to any shorter arms. This can be addressed by playing every active arm in X_m until round $n_m = \left\lceil \frac{2\gamma \log(T(K+1)\tilde{\Delta}_m^2)}{\tilde{\Delta}_m^2} \right\rceil$ in each stage m. The algorithm applies to any time associative stochastic bandit problem with arbitrary $l_t(x_t, a_t)$. By applying a similar argument as in the proof of Theorem 1, the algorithm achieves a regret bound of $\sum_{i=1}^{K} \frac{32\gamma \log\left(T(K+1)\frac{\Delta_i^2}{4}\right)}{\Delta_i} + \sum_{i:\Delta_i>0} \left(\Delta_i + \frac{48}{\Delta_i}\right)$.