# Internship Report:
# Developing Risk-Aware Sequential Decision-Making for Agricultural Decisions under Environmental Risks

**Thomas Michel**

Research internship supervised by **Debabrota Basu** and **Odalric-Ambrym Maillard**

École Normale Supérieure Paris-Saclay – Diplôme ARIA
Équipe Scool, Centre Inria de l'Université de Lille

January 15, 2024

## Contents

# 1 Introduction

Reinforcement learning (RL) is a subfield of machine learning that specifically addresses the development of algorithms and techniques for training intelligent agents to make sequential decisions. It draws inspiration from the concept of learning through interactions with an environment, where the agent receives feedback in the form of rewards or punishments based on its actions. By employing trial-and-error learning, reinforcement learning empowers agents to acquire optimal strategies that maximize rewards, thus finding utility in various domains such as robotics, game playing, and resource management. Nonetheless, conventional RL approaches often neglect the crucial aspect of risk, focusing solely on maximizing expected rewards. This disregard for risk can result in suboptimal decisions in real-world applications where risk-sensitive behavior is of utmost importance.

In recent years, there has been a growing recognition of the need to incorporate risk considerations into RL algorithms. Risk-sensitive reinforcement learning (RSRL) aims to address this gap by enabling agents to make decisions that account for the potential consequences and uncertainties associated with their actions. By explicitly considering risk, intelligent agents can navigate uncertain and risky environments more effectively, leading to improved performance and robust decision-making.

In the context of agriculture and forest management, our application focus in this study, RSRL can play a vital role in enabling farmers and agricultural systems to make more informed and risk-aware decisions. Agriculture is inherently subject to numerous sources of risk, including weather variability, pests and diseases, market fluctuations, and resource constraints. The consequences of poor decision-making in agriculture can have significant impacts on crop yields, livestock health, financial stability, and even food security at larger scales. By accounting for risk, farmers can make more robust and adaptive decisions that align with their risk tolerance and long-term sustainability objectives. By quantifying and optimizing for both immediate rewards and long-term risks, RSRL algorithms can guide farmers towards more sustainable and resilient agricultural practices.

The notion of risk in sequential decision-making can be studied under multiple different perspectives. The first perspective centers on the agent's perception of risk, recognizing that different agents may exhibit distinct risk-preferences when confronted with similar situations. We acknowledge that risk is subjective, and an agent's perception of risk heavily influences their decision-making process. By examining various risk profiles and preferences, we can gain a deeper understanding of how individuals assess and respond to uncertain circumstances. The second perspective delves into the inherent risks within the system itself. As the system operates, it is subject to changes, potential model misspecifications, and even corrupted input signals. These risks arise externally, often beyond the direct control of the decision maker. Factors such as environmental fluctuations, technological disruptions, or unexpected events can introduce uncertainties and impact the reliability and accuracy of the decision-making process. By comprehensively studying these sources of risk, we can develop strategies to mitigate their effects and enhance the robustness of the system. Lastly, the third perspective examines the risks generated by the actions taken within the system. It acknowledges that certain actions may alter the system's dynamics and steer it towards states that are inherently riskier than others. These risks emerge from within the decision-making process itself, as the consequences of actions cascade and potentially introduce additional uncertainties. By scrutinizing the impact of various actions on the system's risk profile, we can identify optimal decision paths that minimize potential risks or maximize rewards while considering the inherent uncertainties at play.

This study primarily focuses on investigating the influence of external risks, dynamics, or events that occur independently of the agent's actions on the optimal strategy. We examine both reward optimization objectives and risk-sensitive objectives. By applying this analysis to forest resource management, we further explore how the optimal strategy, based on local context and information, can adapt in the face of hazards impacting the environment at a broader scale.

The presented contribution has two aspects. Firstly, we develop and analyze a mathematical model of forest growth that incorporates interactions between trees. The novelty of this model lies in its tree-level operation, as opposed to the traditional stand-level approach, allowing for simulation of interactions between neighboring trees. Additionally, the model incorporates multiple environmental hazards with distinct dynamics. We provide a concise investigation into the impact of these external risks on expertise or policies. Furthermore, we have implemented the model in such a way that it can be used independently of this study, adhering to the standard programming interface for reinforcement learning simulated environments.

The second contribution expands upon the forest management case study and proposes a more comprehensive framework for reinforcement learning that accounts for external hazards. These hazards are events which occur independently of the agent's actions and temporarily modify the model's dynamics. We delve into the application of this framework to stochastic multi-armed bandits, offering an algorithm tailored to this context and providing initial elements for its analysis. We conclude by conducting an empirical study to validate our findings and intuitions pertaining to the further pursuit of the work started here.

# 2 Related works

## 2.1 Risk-sensitive approach to agricultural decision-making

This study focuses on analyzing decision-making risks in the context of forest management. Previous research has examined the optimal management of forest resources, specifically addressing risks associated with windthrow and storms. Couture et al. [2016] conducted a study on the management of an uneven-aged forest under the risk of windthrow, utilizing a Markov decision process (MDP) approach. MDP is a widely-used framework for modeling sequential decision-making problems, where an agent makes choices in an environment, and the outcomes depend on the current state and chosen action. At each step, the agent observes the current state, takes an action, receives a reward from the environment, and transitions to a new state according to a probability distribution dependent only on the current state and chosen action. The agent's objective is to identify optimal strategies that maximize rewards over time. The model proposed by Couture et al. [2016] incorporated risk aversion and aimed to minimize the expected discounted windthrow costs over time, resulting in an optimal management strategy tailored to the risk preferences of the forest manager. Couture et al. [2021] further extended the MDP framework to incorporate multiple objectives in sequential forest management under risk. They proposed the use of the Markov Decision Process-Pareto Frontier approach, which allows for the consideration of trade-offs between conflicting objectives. By constructing a Pareto frontier, they identified a range of optimal management strategies that strike a balance between risk reduction and economic returns.

Loisel [2014] conducted a study investigating the influence of storm risk on the Faustmann rotation, a widely employed concept in forestry for determining the optimal period for harvesting trees. Their research aimed to assess the effect of storm risk on rotation decisions and emphasized the importance of integrating risk factors into forest management strategies. In a more recent investigation, Loisel et al. [2022] delved into the role of ambiguity and the value of information in forest rotation decisions when faced with storm risk. The authors incorporated decision-making under uncertainty into the framework of forest rotation modeling and examined how uncertain information impacts optimal management strategies. Their findings highlighted the significance of considering both risk and ambiguity in the decision-making processes related to forest management.

In contrast to the aforementioned studies that focus on computing optimal policies for known models, our study aims to address the challenge of learning within an initially unknown environment by employing a policy that integrates risk considerations. Furthermore, our specific focus is on a tree-level modeling approach that incorporates interactions between trees, allowing us to examine the broader-scale impact of these interactions. To the best of our knowledge, this particular approach has not been explored in the existing literature.

While the previous references concentrate on forest management, Gautron et al. [2022] addressed the issue of assisting farmers in identifying optimal crop management strategies. Their study aimed to develop an effective and risk-aware approach to support decision-making among farmers, incorporating risk considerations into the decision-making process to optimize crop management strategies that minimize potential losses and maximize agricultural productivity.

Although this study shares similarities with our research in terms of integrating risk into decision-making, we adopt a different approach in the context of forest management. Rather than predefining different planning strategies and considering complex outcome distributions, we allow the agent to take instantaneous actions at any stage of forest development.

## 2.2 Risk-sensitive reinforcement learning

Risk-sensitive reinforcement learning (RSRL) is an important area of research, particularly in the context of agricultural sequential decision making, where a risk-neutral strategy is insufficient. Its primary objective is to address the tradeoff between optimizing rewards and ensuring safety, as well as the need for robust decision-making processes.

In our study, we placed emphasis on the modeling of risks and devised a framework that treats risks as events independent of the agent's actions. The primary goal of our work is to maximize the agent's overall reward, even in the face of unfavorable outcomes resulting from uncertain events or environmental hazards. This concept, although related, differs from the conventional notion of risk as applied in risk-sensitive reinforcement learning. Specifically, risk-sensitive reinforcement learning typically considers risk as an inherent characteristic of decision-making in a *stochastic* environment, as opposed to the realization of an undesirable random event as we discuss in this study.

Over the years, multiple approaches to RSRL have been explored. For a comprehensive survey of the field, refer to Garcıa and Fernández [2015]. Additionally, Tan et al. [2022] provide a more specific survey on applications of risk-sensitive multi-armed bandit problems, which we will focus on in the second part of our study.

In this section, we will present an approach to risk-sensitive reinforcement learning (RSRL) that revolves around optimizing risk measures. Although not directly included in the current study, this approach holds potential for future extensions of our work, particularly when incorporating a risk-averse component and accounting for user preferences during the learning process.

In risk-sensitive reinforcement learning, one common strategy is to modify the objective function to optimize a risk measure associated with the distribution of rewards. Various risk measures have been explored in the literature, including variance, VaR (Value at Risk), and CVaR (Conditional Value at Risk).

VaR (Value at Risk) is a risk measure that quantifies the maximum potential loss within a specified confidence level, providing a single value for the worst-case loss. In contrast, CVaR (Conditional Value at Risk) expands on VaR by taking into account average or expected losses beyond the VaR threshold. This comprehensive measure considers the severity of extreme events or unfavorable scenarios, offering a more complete understanding of potential extreme losses.

Several studies have incorporated the CVaR risk-adverse criterion [Rockafellar et al., 2000] as a modification of the objective in the Markov decision problem setting [Tamar et al., 2015, Chow et al., 2015] or the multi-armed bandit setting [Baudry et al., 2021]. Furthermore, some studies have explored a broader range of *coherent* risk measures, with CVaR being one such measure [Maillard, 2013, Saux and Maillard, 2023, Mihatsch and Neuneier, 2002].

One challenge in optimizing risk measures is that they are defined with respect to the distribution of rewards. In contrast, in average/total reward optimization settings, it is often sufficient to estimate the mean reward of each arm without considering the entire distribution.

The risk-sensitive approach, while beneficial in quantifying and managing risks, does come with certain drawbacks. Firstly, it is important to note that the objective policy cannot generally be risk-averse and simultaneously achieve optimal reward. There is a trade-off between risk aversion and maximizing rewards. Additionally, the introduction of risk-sensitive objectives also impacts the learning phase of the algorithm. In general, a learner needs to explore extensively to ensure that it does not miss out on potentially higher rewards. However, such exploration carries inherent risks, which should be minimized in a risk-sensitive setting. For a study on this phenomenon in the context of multi-armed bandits, please refer to Galichet et al. [2013].

Numerous other techniques aimed at incorporating risk-aversion have been extensively studied in the literature. Although not specifically addressed in the current study, these techniques share a common objective of developing safe and robust decision-making strategies. Such approaches would be particularly valuable in complex environments where the associated risks can lead to significant costs, as is the case in domains like forest management.

## 2.3 Related bandit settings

The second part of our work involves developing a reinforcement learning framework that accounts for external risks. In this study, we mainly focus on its implications in the stochastic multi-armed bandit setting. See Section 4.2 for an introduction to the multi-armed bandit problem, and Section 4.3 for the modification we introduced to the basic setting in order to include external risks. Our framework incorporates two types of rewards for each action: a nominal outcome and a hazard outcome. This approach bears similarities to the work introduced by Basu et al. [2022] in the context of bandits. The authors focused on the corrupted bandit problem, where unknown reward distributions in a stochastic multi-armed bandit setting are heavy-tailed and subject to corruption by a history-independent adversary. However, there are notable differences between the problem we address and the one studied by Basu et al. [2022]. In our case, our objective is not solely to find the optimal arm despite the corrupted values. Instead, we strive to identify an arm that effectively combines both the nominal distribution and the hazard (corrupted) distribution. As our primary focus is not the robust estimation of rewards, we also consider different assumptions that result in reward distributions with improved properties. By doing so, we aim to facilitate the analysis process and enhance the overall understanding of the problem at hand.

Another desired aspect of our framework is its capability to handle non-stationarity, which refers to the ability to accommodate changes in problem parameters over time. This has been a long-standing challenge in the field of reinforcement learning, and we specifically try to address it in the context of the multi-armed bandit setting. Several authors have explored this issue across different scenarios, presenting various approaches and methodologies to tackle the challenge of non-stationarity. For instance, Garivier and Moulines [2008] investigated the application of upper-confidence bound (UCB) policies to non-stationary bandit problems with abrupt changes Garivier and Moulines [2008]. A main aspect of their proposed method is to allow the algorithm to forget old observations and continuously seek for new ones in order to keep track of the changes of the model. Slivkins (2008) addressed the Brownian restless bandit problem, where reward distributions evolve continuously over time following a Brownian motion, with small changes occurring from one time step to another Slivkins and Upfal [2008]. The focus was on adapting bandit policies to handle this continuous evolution of reward distributions.

In our specific setting, we consider rewards that originate from either a nominal distribution or a "hazard" distribution, with only the probability of sampling from each distribution changing over time. This characteristic allows us to develop estimates for each arm that remain relevant despite the non-stationarity of the problem, which would be difficult in the case of the papers described above.

Other researches in the field has explored the utilization of mixture models, such as the work conducted by Urteaga and Wiggins [2018]. In their study, the authors utilized Gaussian mixture models to handle the uncertainty associated with the reward distribution of each arm. This approach is also sometimes employed in contextual bandits to approximate the reward distribution based on the available context. However, our proposal here differs slightly in that we model the occurrence of different events as samples from distinct distributions, effectively sampling from the mixture distribution that combines them. In our case, the use of a mixture distribution is inherent to the modeling of risk and not merely a proxy to accommodate the estimation of complex distributions.

## 3 Tree-level model of forest growth with interactions between trees

Forest ecosystems are intricate and interconnected systems, characterized by complex interactions among trees, plants, fungi, and animals. However, our current understanding of these interactions remains incomplete. The prevailing models used in forest management oversimplify the reality by disregarding tree interactions, employing discrete representations, and relying on average behaviors. Moreover, they fail to incorporate localized strategies and nuanced risk dynamics. These limitations predominantly stem from computational constraints, which hinder our ability to devise precise management techniques that account for the unique characteristics of individual trees.

To address these issues, our research endeavors to develop a tree-level model that explicitly considers the interactions between individual trees within a forest. By treating each tree as an autonomous agent, we aim to elucidate the influence of local interactions on optimal management strategies. This approach represents a significant step towards bridging the gap between existing models and the intricate reality of forest ecosystems, thereby facilitating more effective forest management. However, we acknowledge that this approach is still subject to the same computational constraints that have constrained previous models. Consequently, we are compelled to limit the scale of the simulation in terms of the number of trees and, in some cases, the number of policies under consideration.

Furthermore, we explore the applicability of current techniques in managing collective risks associated with the global condition of the forest. By considering individual interactions and group risks holistically, we enhance the comprehensiveness and practicality of forest management strategies.

In the ensuing sections, we will provide an overview of our tree-level model, explicate the methodologies employed, and present our preliminary findings. Extensions upon this basic model are discussed in Section 3.5.

## 3.1 Presentation of the forest model without environmental risks

Let us consider a forest of $n$ trees which interact with their neighbors. The *neighborhood relationship* is described as a graph $G$ with $n$ vertex and in which the edges represent the influence of trees on each other. Here we assume that $G$ is a simple undirected and unweighted graph, but future studies may relax these assumptions without much change in the model. The state of the forest at time $t$ is defined by the height of each of its constituent trees and denoted as the vector $s_t = (s_t^1, ..., s_t^n, H_t)^\intercal \in \mathbb{R}^{n+1}$. We will also later consider an action $a_t$ chosen by the forest manager. We model the growth of the forest as a discrete time linear system of the form

$$s_{t+1} = As_t + Ba_t$$

.

$A$ is the transition matrix of the system, which depends on the graph $G$ of interactions between the trees, as well as two additional parameters, $\alpha$ and $\beta$. $\alpha$ can be thought of as a growth parameter that influences the rate of tree growth, while $\beta$ is an interaction parameter that dictates the strength of interaction between the trees. In this case, the main interaction manifests as a bonus or malus to the growth rate based on the relative size of neighboring trees.

**Dynamic of the model** We denote as $\mathcal{V}_i$ the set of vertices connected to $i$ in $G$. The dynamic of the system without exterior action will be the following

$$\forall i \in [\![1, n]\!], s_{t+1}^i = s_t^i + \alpha(H_t - s_t^i) + \frac{\beta}{|\mathcal{V}_i|} \sum_{j \in \mathcal{V}_i} (s_t^i - s_t^j),$$
$$H_{t+1} = H_t.$$

We can observe that, according to this definition, the last coefficient of the state vector remains constant and is equal to its initial value, denoted as $H$. In this model, tree growth is facilitated by the term $\alpha(H_t - s_t^i)$, which promotes the growth of the tree up to an asymptotic value $H$, while penalizing excessive growth. The last term, $\frac{\beta}{|\mathcal{V}i|} \sum_{j \in \mathcal{V}i} (s_t^i - s_t^j)$, represents the interaction between the trees. Having tall neighbors is detrimental to growth, as they partially block sunlight or absorb more nutrients from the ground. Conversely, small neighbors allow for rapid growth due to the lack of strong competitor.

We can now define the corresponding transition matrix $A$. For all $i \in [\![1, n]\!]$ and $j \in [\![1, n]\!]$,

$$A_{i,j} = \begin{cases} 1 - \alpha + \beta\,\mathbb{I}\{\mathcal{V}_i \neq \emptyset\} & \text{if } i = j \\ -\frac{\beta}{|\mathcal{V}_i|} & \text{if } j \in \mathcal{V}_i \\ 0 & \text{otherwise.} \end{cases}$$

In addition $\forall j \in [\![A, n]\!], A_{j,n+1} = \alpha$, $A_{n+1,j} = 0$ and $A_{n+1,n+1} = 1$.

For instance, with two trees in interaction, we obtain

$$s_{t+1} = \begin{pmatrix} 1-\alpha+\beta & -\beta & \alpha \\ -\beta & 1-\alpha+\beta & \alpha \\ 0 & 0 & 1 \end{pmatrix} s_t + Ba_t.$$

**Actions**

For each tree described by the model, the agent has the option to either allow it to grow or cut it down and directly plant a new one. It is possible to cut multiple trees simultaneously. The harvest action occurs before the growth step and the occurrence of storms.

$$s_{t+1} = As_t + Ba_t.$$

We define $B = -A \times \begin{pmatrix} I_n \\ 0...0 \end{pmatrix}$ and restrict the actions $a_t \in \{K_t s_t | K_t = \text{diag}(a_t^1, ..., a_t^n), (a_t^1, ..., a_t^n) \in \{0, 1\}^n\}$.

With this definition, at step $t$ and for each tree $i$ the agent can either choose $a_t^i = 0$ to let the tree grow or $a_t^i = 1$ to harvest and replant the tree.

**Rewards**

We proceed to define the reward function for the forest manager, which serves as the primary objective to optimize. We have chosen to employ a quadratic reward function, as it is a well-studied objective in the domains of reinforcement learning and optimal control. Furthermore, a quadratic reward function aligns with the rationale of studying rewards proportional selling price of the lumber, as it directly corresponds to the basal area of the tree, which determines the selling price of the lumber. Specifically, we consider a quadratic reward function at time $t$ of the following form:

$$r_t = s_t^\intercal Q s_t + a_t^\intercal R a_t.$$

Here, $a_t^\intercal R a_t$ corresponds to receiving a reward that is quadratic in the size of the harvested trees. The first term, $s_t^\intercal Q s_t$, although not utilized in the experiments, can be used to model the value associated with other ecosystem services that are dependent on having a mature forest, such as carbon storage, biodiversity, or recreational purposes.

**Simulation of forest growth without intervention from the forest manager**

The growth of trees follows a natural pattern characterized by an initial fast growth phase, which gradually slows down over time, eventually reaching an asymptotic value denoted as $H$. This growth process is influenced by the presence of neighboring trees. Specifically, smaller trees experience a reduced growth rate when surrounded by taller neighbors. This phenomenon can be attributed to taller trees obstructing the direct access of sunlight for smaller trees, while benefiting from their own unimpeded access to light. Previous studies have shown that the growth of various tree species is positively correlated with light availability [Rüger et al., 2011].

The specific relationship between trees can lead to convergence towards different asymptotic heights. For instance, Figure 1 illustrates the final heights in a forest where trees are arranged in a grid-like pattern and only interact with their four closest neighbors. In this case, the system stabilizes with trees alternating between two distinct sizes, as taller trees impede the further growth of smaller ones. We should note that the positions of taller and smaller trees, as well as the asymptotic heights of both groups, depend on the initial conditions. Additionally, different relationships between trees can result in diverse patterns or behaviors for the asymptotic height. As demonstrated in Figure 1, when trees in a grid pattern interact with their eight nearest neighbors instead of four, the resulting height pattern is different and the asymptotic values of each group are the same, or very close to each others.

The system remains stable as long as the spectral radius of the transition matrix is less than or equal to 1, and the initial values fall within a plausible range (non-negative heights and trees that are not excessively tall compared to the parameter $H$). For example, when considering a complete interaction graph where all trees interact equally with each other, the following conditions must be satisfied:

$$\begin{aligned} &0 \leqslant \alpha \leqslant 1 \\ &0 \leqslant \beta \leqslant \frac{n-1}{n}\alpha. \end{aligned}$$

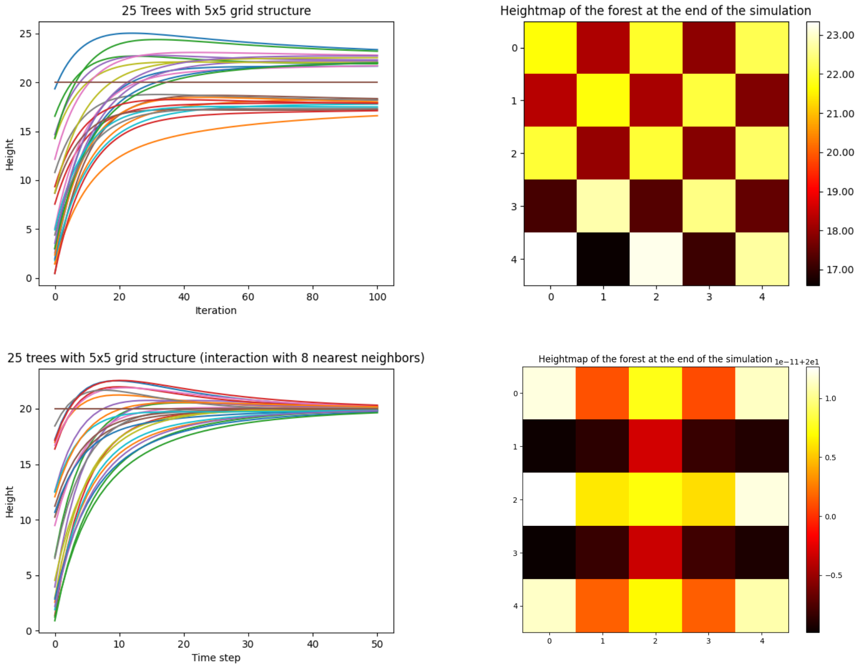


Figure 1: Evolution of tree heights and map of heights at the end of the simulation. Top: Interaction with the 4 closest neighboring trees, Bottom: Interaction with the 8 closest neighboring trees

## 3.2 Comparison of policies on the model without environmental risk

### Description of the considered policies

Two expert policies are compared to a third policy obtained through reinforcement learning using the Proximal Policy Optimization (PPO) algorithm. The first expert policy, referred to as the "cutting age policy", involves setting a fixed cutting age for all trees, which represents the number of steps between planting a tree and harvesting it. There are two versions of this policy: one where all trees are cut simultaneously when their cutting age is reached, and another where each tree's cutting date is offset by a few steps (for implementation details, refer to Section 3.2). It's worth noting that the reward obtained using this strategy may depend on the timing of the first cutting of the trees.

The second expert policy considered in the deterministic setting is the "threshold policy." This policy dictates that a tree should be cut as soon as its height reaches a specified threshold.

For the reinforcement learning approach, an agent is trained using the PPO algorithm introduced by Schulman et al. [2017]. Following training, it is observed that the policy converges to a periodic pattern in which some trees are cut early, while others are allowed to grow for a longer period before being harvested. Figure 3 illustrates the frequency of tree harvests for a grid-aligned forest with interactions among the four nearest neighbors. The figure depicts an alternating pattern of frequently harvested trees and trees allowed to grow further before being harvested. This strategy prioritizes maximizing the overall reward by exploiting the growth potential of larger trees, taking into account the positive impact of having smaller neighbors on tree growth.

The policies can be categorized into two groups based on their granularity. Stand-level policies involve applying the same action to all trees within a stand, such as the cutting age policy, where all trees are harvested simultaneously once the cutting age threshold is reached. Tree-level policies, on the other hand, take personalized actions for each tree based on its current state and neighborhood. The policy learned through the PPO algorithm and the threshold policy both fall into the tree-level category.

**Results**

Table 1 presents the total rewards achieved by the different policies described earlier. The parameters for the threshold policy and the cutting age policies are selected to maximize the total reward (refer to Figure 2). The evolution of the system is deterministic as both the dynamics and policies are deterministic. However, the initial conditions are randomly selected, resulting in slight variability in the total rewards. This small variability explains why the error bands are barely visible.

| | Total reward |
|---|---|
| Threshold policy | 249.61 |
| Cutting age policy | 226.88 |
| Cutting age policy (with offset) | 243.84 |
| Policy learned using PPO | 257.11 |

Table 1: Comparison of the different policies over 100 simulation steps in the deterministic setting (using the best parameters)

The experiments were conducted using the parameters $\alpha = 0.2$, $\beta = 0.1$ and $H = 20$. The total reward is computed over 100 steps of simulation, which amount to multiple cycles of harvest and growth for each strategy.

The results show that the policy learned using the PPO algorithm outperforms the expert policies in the deterministic case. This policy leverages the dynamics of tree interactions to optimize its total reward. The two cutting age-based policies perform similarly, with a slight advantage observed for the policy that offsets the cutting of neighboring trees.

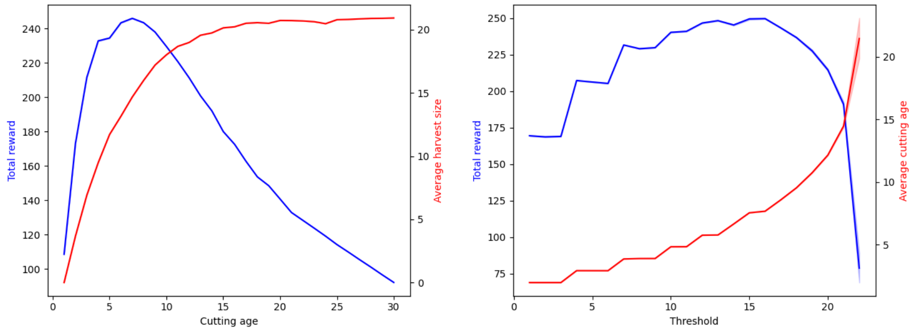


Figure 2: Optimization of the parameters of the simple policies

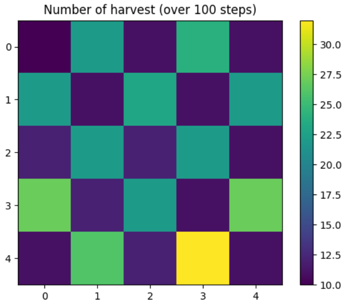


Figure 3: PPO cutting frequencies

## 3.3 Forest model under environmental risks

### Modeling environmental hazards

In this study, we introduce the incorporation of risks, which manifest as stochastic tree destruction events. These events resemble tree cutting actions but occur randomly and are beyond the control of the forest manager. Importantly, the agent does not receive any additional reward for the fallen tree when these events occur in the specific model implementation described below. Losing trees in this manner signifies a net loss for the forest manager.

To model these destructive events, we treat them as if they were actions of a second player by introducing an additional stochastic term into the model:

$$s_{t+1} = As_t + Ba_t + Cb_t. \tag{1}$$

Here, $C$ represents the matrix that defines the behavior of the destructive events, and $b_t$ represents the corresponding actions. The definition of $C$ and the admissible values of $b_t$ are similar to that of matrix $B$ and action $a_t$. The key distinction is that the actions taken by the forest manager, represented by $a_t$, must be taken into account since they are assumed to occur before the destruction of trees by an environmental hazard. Therefore, we include $a_t$ in the definition of $b_t$.

$$C = B = -A \times \begin{pmatrix} I_n \\ 0...0 \end{pmatrix}$$
$$b_t = L_t(s_t - \begin{pmatrix} I_n \\ 0...0 \end{pmatrix} a_t).$$

These definitions allow writing the model dynamics in a simplified manner.

$$s_{t+1} = A(I - L'_t)(I - K'_t)s_t. \tag{2}$$

where $K'_t = \begin{pmatrix} I_n \\ 0...0 \end{pmatrix} K_t$ is controlled by the agent and $L'_t = \begin{pmatrix} I_n \\ 0...0 \end{pmatrix} L_t$ is controlled by the environment.

The modeling of an environmental hazard primarily revolves around the selection of $L_t$, which abides by the same constraints as $K_t$, the matrix that determines which trees are to be cut by the agent. In the following paragraphs, we introduce two approaches for defining the random process that generates $L_t$ to model storms and forest fires, respectively.

**Dynamic of storms**

We will present a straightforward model of a storm, which occurs randomly and independently of the agent's actions. At each time step, there is a predetermined probability denoted as $p_{storm}$ that a storm transpires in the forest. During a storm event, each tree has the potential to be individually destroyed, with a windthrow probability influenced by the heights of its neighboring trees. The probability distribution is intentionally designed to mimic the effects of wind shielding, although it should be noted that the model itself may not accurately reflect reality. When a central tree is surrounded by taller trees, it benefits from protection against windthrows. Conversely, a tall tree with few or no neighbors is highly susceptible to such risks.

The specific probability distribution takes into account these factors to determine the likelihood of windthrow for each tree during a storm event.

To complete the general model of risk presented earlier, we define $L_t = Y_t^{storm} \text{diag}(Z_t^1, ..., Z_t^n)$, with $Y_{storm} \sim \mathcal{B}(p_{storm})$ and $Z_t^i \sim \mathcal{B}(p_t^i)$, where

$$p_t^i = \exp\left( - \sum_{j \in \mathcal{V}_i} s_t(i)/(HD) \right). \tag{3}$$

The parameter $D$ represents the destructive power of the storms experienced by the forest. As $D$ increases, the probability of a tree being damaged during a storm also increases.

In this study, we have chosen to reduce the variability of the storms by fixing the parameters of the risk model, rather than sampling them from a specific distribution at each time step. This approach provides a consistent and controlled framework for analyzing the effects of the storms on the forest.

In this study, we have opted to minimize the variability of storms by fixing the parameters of the risk model, instead of sampling them from a specific distribution at each time step. This approach offers a consistent and controlled framework for investigating the influences of storms on the forest.

An interesting expansion of the model would involve permitting both the values of $D$ and $p_{storm}$ to vary over time, thus incorporating factors such as climate changes. This would facilitate a more dynamic depiction of the evolving environmental conditions and their ramifications for the forest.

**Dynamic of forest fires**

Forest fire risk can be modeled similarly to the windthrow risk, where it is treated as the action of a non-controllable opposing player cutting trees without the agent receiving any reward. However, the dynamics of tree destruction during a forest fire event are different.

We introduce the variable $L_t = Y_t^{fire}\text{diag}(Z_t^1, ..., Z_t^n)$, where $Y_{storm} \sim \mathcal{B}(p_{storm})$ and $Z_t = F(s_t)$. Here, $F$ is a stochastic process that simulates the propagation of a forest fire and returns a vector indicating the destroyed trees (1 if the tree is destroyed, otherwise 0).

The fire propagates for a predetermined number of rounds, which is independent of the system's time steps (everything occurs within one time step). At each round, we consider every pair of neighboring trees $(i, j)$. If the tree $i$ is burning, then the tree $j$ catches fire with a probability $p_{i,j}$ given by

$$p_{i,j} = \frac{0.5}{1 + e^{(s_t^i - s_t^j)}}.$$

This model of fire propagation, inspired by the work of Morales et al. [2015], is intentionally minimalistic, as it includes only variables and parameters relevant to our model of forest, such as tree heights and the interaction graph. The underlying concept of this definition is to account for the difference in fire dynamics between shrublands and forests. Since each tree has a chance to catch fire from each of its burning neighbors at each round, the risk increases if a tree is surrounded by fire.

## 3.4 Comparison of policies on the model under environmental risk

**Description of the considered policies**

Despite the introduction of heightened risks, the previously established policies remain applicable. However, as the probability of environmental hazards escalates, there is a substantial decrease in average reward, accompanied by an increased variability in total reward, as illustrated in Figure 4. Under such circumstances, it becomes crucial to consider these risks when devising policies.

In particular, we consider an additional expert policy that draws inspiration from an existing strategy employed to combat forest fires, while taking into account the nature of the risks involved. We modify the standard threshold policy by implementing artificial clearings, effectively partitioning the forest into multiple distinct blocks. This approach impedes the propagation of forest fires across these blocks, thereby mitigating the associated risks.

Through the integration of this modified policy, forest managers can strategically create clearings within the forest, enhancing its resilience against fire incidents and minimizing the potential damage caused by the spread of fires.

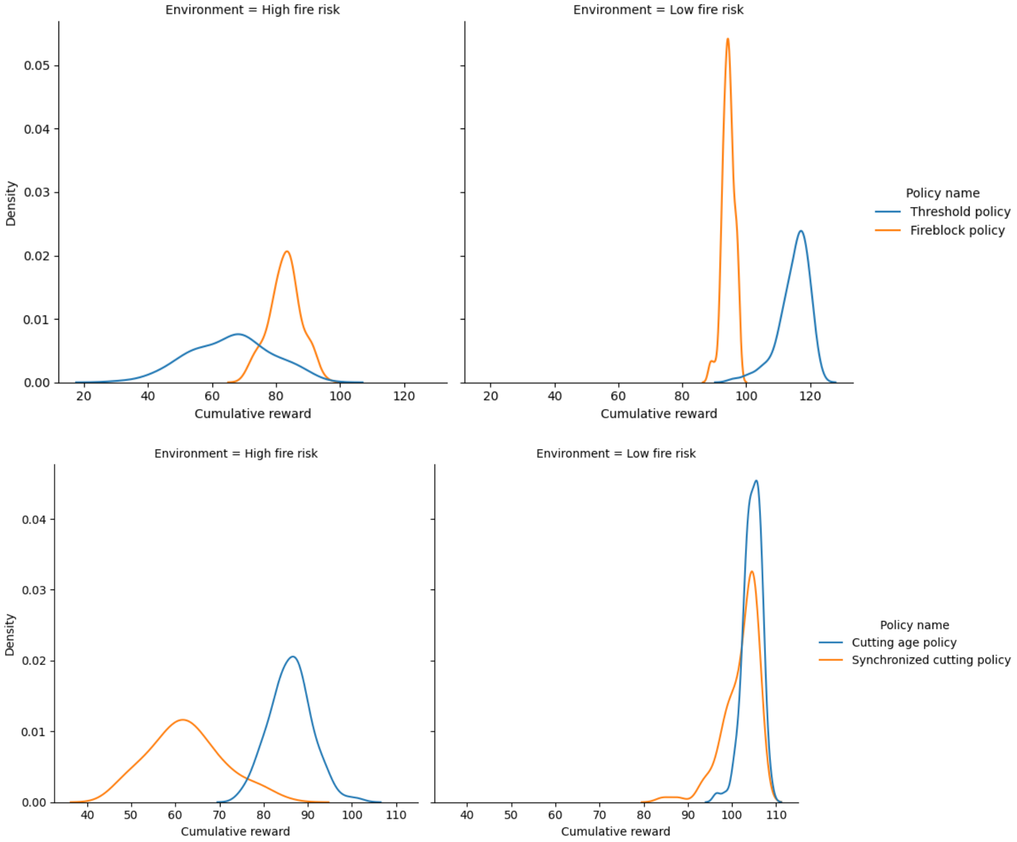
**Results**

The policies are evaluated in two distinct environments characterized by varying probabilities of environmental hazards. In the high-risk environment, the occurrence of a forest fire or storm stands at 20% per time step, whereas the probability diminishes to a mere 1% in the low-risk environment.

The outcomes depicted in Figure 4 employ optimized parameters tailored to each environment, including the optimal threshold level for threshold-based policies. Despite yielding lower overall rewards in the low-risk environment, the implementation of clearings to obstruct fires outperforms the standard threshold policy. Moreover, this approach entails reduced risk as it is less prone to generate significantly meager rewards resulting from successive fires that lead to the loss of the entire forest.

It is worth highlighting that, while both cutting age policies yield similar outcomes in low fire-risk environments, the stand-level policy—comprising the simultaneous removal of all trees within a stand—is more susceptible to forest fires in comparison to the policy that assigns distinct cutting dates to individual trees, thereby creating an unevenly aged forest.

Furthermore, policies acquired through the PPO algorithm exhibit either equivalent performance to the best expert policy, in terms of average reward and risk, when trained and evaluated in environments featuring the same frequency of environmental hazards, as depicted on Figure 5. A policy trained on a forest frequently affected by storms will excel in a similar environment but will prove largely suboptimal if the probability of storms decreases. The same principle applies to policies trained in environments where storms are exceedingly rare. To ensure a fair comparison, only the most favorable version of the learned policy should be juxtaposed with the threshold policy, given that the threshold parameter is selected to best suit the respective environments in both cases.


Environment = High fire risk

Environment = Low fire risk

Environment = High fire risk

Environment = Low fire risk

Figure 4: Distribution of total rewards for a forest affected by fires (KDE plots obtained from empirical distribution over 100 runs)

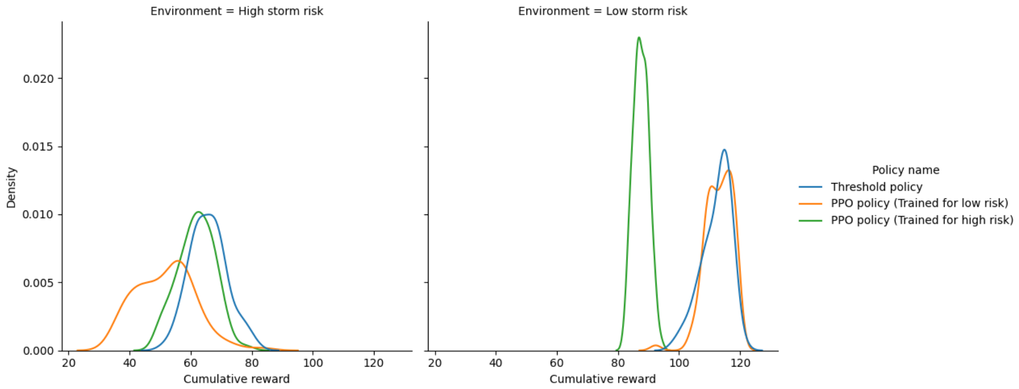


Figure 5: Distribution of total rewards for a forest affected by storms (KDE plots obtained from empirical distribution over 100 runs)

## 3.5 Discussions on the forest model study

The developed model allows for the examination of the effects of external risks on a forest with interconnected tree interactions, albeit with certain limitations. Nonetheless, it provides valuable insights into the impact of different policies. One notable observation is that implementing a global strategy, such as harvesting an entire plot at once, proves to be inefficient in terms of both overall earnings and the vulnerability of the forest to environmental hazards. Despite being recognized as problematic, this strategy is still utilized by forest managers due to cost-saving benefits, as it eliminates the need for individual tree monitoring and the simultaneous management of multiple stands during a single season. Exploring alternative strategies using models like the one proposed here has the potential to foster more sustainable and resilient forest management practices.

Furthermore, the study demonstrates the viability of discovering tree-level strategies through learning techniques like reinforcement learning. However, future investigations should prioritize the application of more risk-sensitive techniques.

Lastly, the study emphasizes the importance of adapting strategies based on both the actual level of risk and the risk preferences of the forest manager. As previously demonstrated, policies that perform well under specific frequencies of environmental hazards may exhibit a substantial decrease in performance compared to their robust and risk-averse counterparts. A sound forest management policy should be capable of performing well despite the discrepancies between the model and reality, which may arise from imperfect modeling or changes in the system dynamics over time.

In future research, it would be worthwhile to investigate variations of the current model that incorporate additional complexities. One potential direction is to consider a forest composed of multiple species of trees, each characterized by unique growth dynamics and interactions. This would provide a more realistic representation of diverse forest ecosystems. Another avenue for exploration is to incorporate varying degrees of interaction by utilizing a weighted interaction graph, where the influence of a neighboring tree is stronger when they are in closer proximity. This would capture the spatial aspect of tree interactions more accurately. Additionally, introducing limitations on visible information and manageable areas, such as observing or managing only a portion of the forest, could offer valuable insights into practical forest management scenarios.

Specifically, the latter variation could prove advantageous for studying learning algorithms in this environment, as it addresses computational challenges stemming from the exponential growth of actions relative to the number of trees. Furthermore, to enhance the manageability of the problem for learning algorithms and facilitate formal study, a discrete model based on the Markov decision process framework can be considered.

It is worth noting that our implementation[1] includes the multi-species forest, area restrictions, and discrete model. However, further exploration and analysis of these aspects are still pending.

The present case study focuses on the examination of external risks in forest management, which necessitates modeling environmental hazards as randomly-occurring events that disrupt the normal dynamics of the environment in a specific manner and for a limited duration. These events transpire independently of the agent's actions and alter the state of the forest based on its state immediately prior to the event. Importantly, the hazard dynamics are not influenced by the historical context of the forest. This characteristic can be advantageous in terms of adapting the learning process and policies if the environment undergoes changes. However, as far as our knowledge extends, this particular type of system has not been explored in the field of reinforcement learning. Therefore, we propose to delve into this topic in the subsequent section.

# 4 Reinforcement learning with external risks

## 4.1 General framework of reinforcement learning with external risks

In order to model the impact of external risks, we introduce a modification to the standard RL setting with random events that occasionally and temporarily modify the dynamics of the system. At each time step, in addition to observing the state/context, the agent can also access a forecast estimating the probability of an event occurring that could disrupt the normal system dynamics. After the agent selects an action, the system will either evolve following its standard dynamics or, in the event of a disruptive occurrence, follow alternative dynamics. The agent then receives a reward and is notified of the event occurrence. Such an event only affects a single time step.

This setting allows for modeling environmental hazards. Let us consider the task of growing crops in a field that can be affected by storms. At each step, the manager receives a weather forecast for the next few weeks and can decide what strategy to adopt based on the current state of the field to minimize losses in the event of a violent storm. The manager can choose to take more drastic preventive actions as the probability of experiencing a storm increases, in order to balance the cost of taking suboptimal actions with the risk of loss due to the external event. The occurrence of a "storm" event happens independently of the actual state of the field, but the outcome of the storm can vary depending on the type of crops that were harvested and those that remained. On larger time scales, one could consider a hazard occurrence model based on both a global trend (e.g., global warming) and seasonal effects (predictions for each month based on previous years) and plan accordingly. Designing a sound strategy becomes more manageable as the normal dynamics and the dynamics associated with extreme events are learned separately and independently. Therefore, if the occurrences of such events increase or decrease, one can combine the two models differently without having to learn the entire model again.

In this preliminary study, our primary focus will be on investigating how this problem can be mapped to learning optimal actions within the context of stochastic multi-armed bandits with external risks.

[1] Gym environment in Python, accessible at https://github.com/Thomick/forest-risk-rl

## 4.2 Technical background on multi-armed bandits

Stochastic multi-armed bandits present a fundamental challenge in sequential decision-making under conditions of uncertainty. Within this framework, an agent is faced with a set of options, referred to as "arms," whose reward distributions are unknown. The agent's objective is to achieve a balance between exploring different arms to gather information and exploiting arms that offer higher expected rewards, aiming to maximize its cumulative reward over time. This problem can be analogized to a gambler positioned in front of a series of slot machines (arms) characterized by diverse probabilities of payout. The gambler's objective is to formulate an optimal strategy to maximize their winnings.

More formally, we define a multi-armed bandit problem involving $K$ arms. In each time step $t$, the agent must select an arm $a \in 1, ..., K$. A reward $r_t$ is subsequently sampled from the unknown distribution associated with the selected arm. The agent observes this reward and continues the process until the trial concludes. The objective is to maximize the expected cumulative reward over the $n$ steps of the trial.

Let us consider the regret of the algorithm over $n$ steps. The regret serves as a metric that quantifies the cost associated with employing a specific strategy in comparison to the optimal one. At each step, we compute the difference between the average reward obtained from the chosen arm and the optimal arm. The regret is obtained by aggregating these differences, also referred to as "gaps," throughout the entire trial. This enables the establishment of an objective measure that evaluates the performance of the algorithm over a specified duration. In the standard multi-armed bandit setting, the regret can be defined as follows:

$$R_n = \sum_{t=1}^{n} \max_a \mu_a - \mathbb{E}\left(\sum_{t=1}^{n} r_t\right),$$

where $\mu_a$ is the true average reward of the distribution associated to the arm $a$.

Multiple strategies have been developed to address the multi-armed bandit problem. In this study, we will focus on a specific paradigm known as "Optimism in the face of uncertainty." Optimism in the face of uncertainty is a fundamental concept in reinforcement learning, proposing that an agent should adopt an optimistic approach when exploring uncertain options in order to gather more information and potentially discover more rewarding outcomes. A popular algorithm that embodies this concept is UCB1 (Upper Confidence Bound) [Auer et al., 2002]. UCB1 achieves a balance between exploration and exploitation by assigning optimistic values to unexplored actions, thereby encouraging the agent to explore them. As the agent accumulates more information over time, the level of optimism diminishes, leading to a refined decision-making process based on acquired knowledge.

The authors establish an optimistic index for each arm, which serves as an upper bound of a confidence interval for the average reward associated with that arm. This interval represents the range within which the true average reward of the arm is likely to fall with high probability. The upper confidence bound for arm $i$ at time $t$ ($\mathrm{UCB}_i(t)$), given a confidence parameter $\delta$, can be defined as follows:

$$\mathrm{UCB}_i(t, \delta) := \hat{\mu}_i(t) + \sqrt{\frac{2\log(\frac{1}{\delta})}{T_i(t-1)}}, \tag{4}$$

where $\hat{\mu}_i(t)$ represents the average reward estimated by the algorithm at time $t$, and $T_i(t-1)$ denotes the number of times arm $i$ has been pulled up to time $t-1$. This index ensures that the true average reward of arm $i$ is below $\mathrm{UCB}_i(t)$ with a probability of at least $1-\delta$. We can now present a variant of the UCB algorithm for a fixed confidence parameter $\delta$, which differs from the UCB1 algorithm [Auer et al., 2002] in that it requires predefining the confidence level, typically as a function of the total number of steps in the trial (horizon), while UCB1 is independent of the horizon.

**Algorithm 1** UCB($\delta$)

```
Require: Number of arms $K$, Horizon $T$
for time steps $t = 1, \dots, T$ do
    Choose action $i \in \operatorname{argmax}_{j \in [\![1,K]\!]} \mathrm{UCB}_j(t, \delta)$
    Observe the reward $r_t$
    Update the estimates and confidence bounds
end for
```

Given a finite horizon $n$, we can establish guarantees on the average performance of the algorithm, as stated in Proposition 1.

**Proposition 1** (Regret bound for UCB($\delta$))**.** *Consider* UCB($\delta$) *as shown in Algorithm 1 on a stochastic $k$-armed 1-subgaussian bandit problem. For any horizon $n$, if $\delta = 1/n^2$, then*

$$R_n \leqslant 3\sum_{i=1}^{k} \Delta_i + \sum_{i:\Delta_i>0} \frac{16\log(n)}{\Delta_i}.$$

A more comprehensive description of UCB($\delta$) and the proof for Proposition 1 can be found in Chapter 7 of the book by Lattimore and Szepesvári titled "Bandit Algorithms" [Lattimore and Szepesvári, 2020].

## 4.3 Problem statement: Bandits with external risks

In this section, we extend the conventional multi-armed bandit setting to incorporate external risks. We consider a multi-armed bandit problem featuring $K$ arms. Each action $a \in 1, \dots, K$ is associated with two distinct reward distributions, denoted as $P_a$ and $Q_a$. During each round $t$, the learner selects an action $a \in 1, \dots, K$ and receives a random reward $r_t$ drawn from distribution $P_a$ with probability $1 - p_t$, or from distribution $Q_a$ with probability $p_t$. Prior to choosing an arm, the learner observes the probability $p_t$ and subsequently receives information denoted as $o_t$, indicating which distribution was actually used along with the associated reward. Specifically, if $o_t = 0$, it signifies that the reward was drawn from distribution $P_a$, while $o_t = 1$ implies that it was drawn from distribution $Q_a$.

In the context of the finite horizon setting, our objective is to optimize the cumulative reward within a specified time horizon $T$. Let $\mu_a$ denote the expected value of rewards sampled from distribution $P_a$ for each action $a$, and let $\lambda_a$ represent the expected value from distribution $Q_a$. In this setting, we define the regret as follows:

$$R_n = \sum_{t=1}^{n} \max_a((1-p_t)\mu_a + p_t\lambda_a) - \mathbb{E}\left(\sum_{t=1}^{n} r_t\right).$$

As an additional assumption, and to adhere to the idea of external risks, we assume that the probability of an event $p_t$ is independent of the agent's actions prior to time instant $t$. This assumption ensures that the event probability remains unaffected by the agent's choices leading up to a particular time step. However, it is important to note that in the non-stationary case, further assumptions may be required, which will be addressed later.

## 4.4 Study of bandits with stationary event probability

We commence our analysis by considering a scenario in which the probability of selecting from the "hazard" distributions $Q_i$ when pulling arm $i$ remains fixed. In this case, the optimal arm and the regret incurred by selecting a suboptimal arm do not change over time. Consequently, we encounter a classical stochastic multi-armed bandit problem, albeit with the distinction that, contrary to the typical assumptions made regarding the reward distribution (such as boundedness, upper-boundedness, subgaussianity, identical variances, or straightforward Bernoulli distribution), we have a mixture of distributions in this setting. For the purpose of our analysis, we will assume that both $P_i$ and $Q_i$ for each arm $i$ are $\sigma$-subgaussian, as defined below. It is important to note that, for the proposed algorithm and the stationary case, we do not require any assumptions regarding the means of the distributions.

**Definition 1** (Subgaussianity)**.** *A random variable $X$ is said to be $\sigma$-subgaussian if, for all $\lambda \in \mathbb{R}$, the following inequality holds:* $\mathbb{E}[\exp(\lambda X)] \leqslant \exp\left(\lambda^2\sigma^2/2\right)$.

Subgaussianity is a property that is satisfied by several commonly used distributions, including Bernoulli, Gaussian, and bounded distributions. This property, in particular, offers a valuable characteristic that we will exploit to construct an index for our optimism-based algorithm.

**Proposition 2** (Hoeffding bound)**.** *Suppose that the variables $X_i, i = 1, \ldots, n$ are independent, and $X_i$ has mean $\mu_i$ and sub-Gaussian parameter $\sigma_i$. Then for all $t \geqslant 0$, we have*

$$\mathbb{P}\left[\sum_{i=1}^{n} (X_i - \mu_i) \geqslant t\right] \leqslant \exp\left(-\frac{t^2}{2\sum_{i=1}^{n}\sigma_i^2}\right).$$

Proposition 2 is a classical property of sums of subgaussian random variable which proof can be found in [Vershynin, 2018, Theorem 2.6.2] for example.

To address the problem at hand, we present Algorithm 2, denoted as MixUCB (UCB for mixture distributions), which is based on the principle of optimism in the face of uncertainty just like UCB1. However, UCB1 is not suitable for online learning in the context of our problem. The presence of a mixture of subgaussian distributions makes it challenging to handle unknown and varying variances for each arm's reward distribution. While previous studies have explored the case of unknown variances for Gaussian variables [Auer et al., 2002, Cowan et al., 2017], these approaches are not applicable to our setting since a mixture of Gaussians is not itself Gaussian. Instead, we propose a modified version of UCB1 that employs a new index tailored for mixtures of subgaussian variables. This modified algorithm enables separate estimation of the means of the two distributions and provides convenient bounds for analysis. We believe that this algorithm holds promise even in the non-stationary case when the risk occurrence probability $p_t$ is directly provided to the algorithm.

We denote the number of pull of arm $i$ at step $t$ as $T_i(t)$ and the number of such pulls that resulted in a reward sampled from $P_i$ as $T_i^{\mu}(t)$ (resp $T_i^{\lambda}(t)$ for the ones drawn from $Q_i$).

For each arm $i$, let us consider an upper confidence bound similar to the one suggested by Auer et al. [2002].

$$\begin{aligned} \mathrm{UCB}_i(t,\delta,p) &:= (1-p)\hat{\mu}_i(t) + p\hat{\lambda}_i(t) + \beta_i(t,\delta,p), \\ \beta_i(t,\delta,p) &:= \sqrt{2\log\left(\frac{1}{\delta}\right)\left(\frac{(1-p)^2}{T_i^{\mu}(t-1)} + \frac{p^2}{T_i^{\lambda}(t-1)}\right)}. \end{aligned} \tag{5}$$

By convention, we will assign $\beta_i(t)$ to be $+\infty$ if either $T_i^{\mu}(t-1)$ or $T_i^{\lambda}(t-1)$ is zero and their associated probability is strictly positive (i.e. one of the two distribution was not sampled but has chances of being sampled).

**Algorithm 2** MixUCB

```
Require: Number of arms $K$, Horizon $T$
for time steps $t = 1, \dots, T$ do
    Observe $p_t$
    Select arm $i \in \underset{j \in [\![1,K]\!]}{\operatorname{argmax}} \mathrm{UCB}_j(t, \delta, p_t)$
    Play the arm and observe the reward $s_t$ and the event occurrence indicator $o_t$
    Update the relevant values
end for
```

**Risk-informed case** Let's first consider the case referred to as the risk-informed case, which corresponds to the setting where the values of $p_t$ are provided to the algorithm.

**Theorem 1.** *If $\forall t, p_t = p \in [0,1]$ and the value of $p$ is known, then for all horizon $n$, Algorithm 2 with $\delta = \frac{1}{n^3}$ satisfies*

$$R_n \leqslant \sum_{i:\Delta_i(p)>0} \left(3 + \frac{1}{\min(p, 1-p)}\right)\Delta_i + \frac{48\log(n)}{\Delta_i(p)}.$$

We can observe that Theorem 1 provides a regret bound similar to that of UCB1. The factor $\frac{1}{\min(p,1-p)}$ arises from the need to sample from both distributions of each arm. This is necessary to avoid significantly underestimating the average reward by solely sampling from the "bad" distribution. However, considering the motivation behind modeling environments affected by external hazards and assuming that the nominal distribution $P_i$ for each arm $i$ has a higher mean than the hazard distribution $Q_i$, a simple modification of the algorithm allows us to replace the aforementioned term with $\frac{1}{1-p}$ while maintaining a sufficiently optimistic index. This new term is reasonable as long as the probability of the hazard is small.

**Risk-oblivious case** In the case where the reward distributions are assumed to be stationary, we can estimate the mixture parameter using the history of event occurrences. We propose a modification of Algorithm 2 to accommodate situations where the agent cannot directly observe the actual value of $p_t$. Instead, we replace $p_t$ with an estimate $\hat{p}_t$ throughout the algorithm, given by the formula:

$$\hat{p}_t = \frac{\sum_{s=1}^{t-1} o_s}{T_i(t-1)}.$$

This version of the problem can be seen as an intermediate step towards the non-stationary setting, as the estimates form a sequence that varies around and ultimately converges towards the actual mixture parameter. The key distinction is that, unlike in the non-stationary setting, the actual gap between the average rewards of the two arms does not change with the value of $\hat{p}_t$.

**Theorem 2.** *If $\forall t, p_t = p \in [0,1]$ and the value of $p$ is unknown, then for all horizon $n$, Algorithm 2 with estimated event probability and $\delta = \frac{1}{n^3}$ satisfies*

$$\limsup_{n\to\infty} \frac{R_n}{\log^3(n)} \leqslant C \tag{6}$$

*where $C$ is a constant that depends only on the considered bandit instance and event parameter.*

**Remark.** *The omission of the constant is due to its complexity and technical challenges. The proof implies that making better choices could potentially result in both enhanced constants and improved asymptotic behavior. The presented result exhibits poly-logarithmic regret, although there is room for further improvement. While there is no corresponding lower bound, in certain specific cases like Bernoulli rewards, standard bounds can be applied with a regret on the order of* $\log(n)$.

## 4.5 Study of bandits with time-varying event probability

In this section, we will delve into the non-stationary scenario, wherein the mixture parameter, denoted as $p_t$, varies to reflect changes in the probability of an event disrupting the normal dynamics or, in the case of bandits, sampling rewards from an alternative distribution. Due to time constraints during the internship, a comprehensive study of Algorithm 2 in the non-stationary case is lacking. However, we will outline preliminary components and key insights that pave the way for a complete analysis.

In contrast to other non-stationary bandit scenarios documented in the literature, where agents must update their internal estimations and discard old information to adapt to model changes, the agent in this setup can independently estimate the two distributions, which remains relevant throughout the trial. Only the mixture parameter undergoes modification, thereby altering the reward distribution over time. The knowledge of the mixture parameter can either be directly given to the agent (risk-informed), obtained through an external model (such as weather forecast utilizing information unavailable to the agent), or estimated dynamically by the agent based on the event history (risk-oblivious). It is crucial to emphasize that the estimation of the mixture parameter remains unaffected by the agent's choices. This is due to the fact that the probability of an event remains consistent regardless of the chosen action, and the rewards associated with the arm do not play a role in the estimation of the event probability.

From this point onward, we present the remarks and insights obtained from the preliminary study of bandit settings with external risks and time-varying event probabilities. Our primary focus is on the risk-informed scenario, where the agent is equipped with the knowledge of event probabilities.

**Arm optimality** The optimal action and the difference between the expected rewards of the best and second-best actions are both contingent on the value of $p_t$. As a result, these quantities can vary from one round to another. However, it is important to note that each arm can be considered optimal within at most one specific range of values for $p_t$. Moreover, the optimality gap, which refers to the difference between the expected rewards of the best and second-best actions, as well as the individual gaps of each action are continuous and characterized by piecewise linear behavior. In addition, it can be easily shown that the gap between each arm and the optimal arm is a convex function of the event probability $p_t$.

**Convergence toward a switching point** As previously mentioned, an arm can be optimal within at most one interval of event probability. As the event probability approaches the edges of this interval, the gap between the best and second-best arms decreases linearly until it reaches zero. At this point, a switch occurs between the best and second-best arms.

According to standard sample complexity lower bounds for bandit problems, the number of samples required to distinguish between two arms is approximately $\frac{1}{\Delta^2}$, where $\Delta$ represents the difference between the mean rewards of the two arms. Considering Gaussian rewards, the confidence interval is typically on the order of $\sqrt{\frac{2\log(n)}{t}}$, where $n$ is the total number of time steps.

Let's consider a specific instance of a bandit with two arms: $(\mu_1, \lambda_1) = (1, 0)$ and $(\mu_2, \lambda_2) = \left(\frac{3}{4}, \frac{1}{4}\right)$. Additionally, we have the sequence $(p_t)_{t=1}^n = \left(\frac{1}{2} + \frac{1}{t+1}\right)_{t=1}^n$. In this case, the gap between the optimal arm and the other arm is $\Delta(t) = \frac{1}{2t}$. It is worth noting that in this specific problem instance, the sub-optimality gap decreases linearly with the number of time steps, which is faster than the shrinking rate of the confidence interval. As a result, there is a possibility of sampling the suboptimal arm, which remains constant throughout the run, a linear number of times.

We propose a conjecture asserting that in scenarios where the sub-optimality gap, denoted as $\Delta(t)$, approaches zero, any reasonable algorithm is expected to sample suboptimal arms a linear number of times, given that $\Delta(t) = \mathcal{O}\left(\frac{1}{\sqrt{t}}\right)$. If this conjecture holds true, it would imply a lower bound on the regret of at least the order of $\sum_{t=1}^{n} \frac{1}{\sqrt{t}} = \mathcal{O}(\sqrt{n})$ for this class of instances. This is due to the linear increase in the number of suboptimal pulls being counterbalanced by the reduction in the sub-optimality gap. Notably, this dependence on the horizon would be higher compared to the case with a stationary event probability, which has a regret order of $\log(n)$. It is important to exercise caution when comparing this conjecture to the lower bound stated by Garivier and Moulines [2008], which asserts that any policy has an expected regret of at least the order of $\sqrt{n}$. The comparison should be approached with care since the settings between the two studies are vastly different, with minimal assumptions on the event probabilities and complementary information in our case, as opposed to abrupt changes in arm reward distributions at unknown time instants and different assumptions on the distributions in the study of Garivier and Moulines [2008]. Therefore, there is no indication that the complexity of the two settings should be similar. Further investigation is necessary to establish an appropriate lower bound for the problem at hand.

To partially support this intuition, we conducted experiments in Section 4.6 specifically with the proposed algorithm, MixUCB.

**Discrepancy between mixture parameter and past history** Another challenge that may arise is the discrepancy between the number of observed samples and the current event probability. In extreme cases, the probability can abruptly transition from being close to zero at the beginning of the trial to nearly 1 towards the end. In such situations, the estimation based on a limited number of samples from the hazard distribution can become highly inaccurate, resulting in a sudden increase in suboptimal pulls. While MixUCB effectively addresses this issue by adjusting the confidence interval accordingly, analyzing and establishing a uniform upper bound on the regret for this case can be complex. Additional assumptions on the variations of the event probability might be required.

In the analysis of the stationary case, this issue was addressed by noting that the probability of encountering a large discrepancy is small, assuming the event probability remains constant. This was achieved by examining the tail probabilities of the random variable that counts the number of pulls from each distribution of an arm.

**Relevant assumptions** Based on the previous observations, we recognize the need to introduce additional assumptions to fully analyze bandit problems with time-varying event probability.

One possible assumption is to limit the rate of change of $p_t$ by assuming that the event probability is Lipschitz continuous with respect to $t$. While this assumption can help control the discrepancy between the proportion of samples from each distribution and the current event probability, its effectiveness is diminished by the fact that we provide the exact event probabilities to the algorithm. Even in the extreme case described earlier with a sudden change in event probability, the situation after the abrupt change is similar to the one at the beginning of the run. No information is lost, and no additional bias is introduced because we do not have an erroneous estimate of the event probability. Consequently, the situation with a sudden change of event probability from 0 to 1 should result in a regret that is no greater than the sum of the regrets from runs in which the event probability remains close to 0 or close to 1.

An important concern highlighted earlier is the convergence of the event probability towards the switching points of the optimal arm. In such cases, algorithms can face difficulties and may result in a regret of the order of $\sqrt{n}$, where $n$ represents the horizon of the trial. While we have not provided a formal proof for this claim, it raises a significant concern.

In general, this linear rate of suboptimal pulls can arise whenever there exists a subsequence of event probabilities that converge towards switching points at a sufficiently fast rate. The specific conditions leading to these cases need to be further studied and investigated. It is important to either rule out these scenarios or consider them as a characterization of the complexity of bandit problems with external risks.

Another interesting approach to studying the problem would be to consider the sequence of event probabilities as samples from a probability distribution, which can either be stationary or evolve over time. This perspective would allow us to define an average regret with respect to the distribution of event probabilities. It is likely that similar issues as those described earlier, such as convergence towards switching points and the linear rate of suboptimal pulls, would still arise in this framework. A careful analysis might shed further light on the necessary assumptions regarding the distribution of the event probability parameter.

### 4.6 Experimental results

In this section, we present the performance evaluation of the algorithms under study[2]. The algorithms being compared are as follows:

- The proposed algorithm for this scenario is called MixUCB (Algorithm 2).
- The baseline algorithm is UCB1 [Auer et al., 2002]. We have arbitrarily chosen to use UCB1 instead of UCB($\delta$), which only differs slightly in the definition of the index but yields similar empirical results in the finite horizon setting.
- The version of MixUCB with probability estimation is designed for the risk-oblivious case, as discussed in Section 4.4. In this version, the agent is not informed about the probability of occurrence of an event. This adaptation is only suitable when the event probability remains constant. It allows for a fairer comparison with UCB1, as it does not receive the additional information of the probability. However, the agent can still observe from which distribution the arm's reward was sampled.
- The risk-informed $\varepsilon$-greedy algorithm is an adaptation of the standard $\varepsilon$-greedy algorithm. It estimates the average rewards of each arm by computing the empirical average of each distribution and combines them according to the provided probability of event $p_t$. Then, it selects the arm with the highest estimate with probability $1 - \varepsilon$, or a random arm with probability $\varepsilon$.
- The OFUL algorithm [Abbasi-Yadkori et al., 2011] is an optimism-based algorithm specifically designed for addressing the stochastic linear bandit problem (for a comprehensive introduction to the problem and detailed analysis of the OFUL algorithm, please refer to Abbasi-Yadkori et al. [2011]). We can transform an instance of the problem of stochastic bandits with external risk into a linear bandit problem, although the common assumptions in the analysis of this setting, such as additive noise dependent on the chosen action and its context, may not hold in this case.

  The transformation process involves associating a $K$-armed bandit problem with a $K$-armed linear bandit problem, where the contexts have a dimension of $2K$. At each time step, if the event probability is denoted as $p_t$, the context of arm $i$ is represented by a $2K$-dimensional vector with zeros in all coefficients except at positions $2i$ and $2i + 1$, which are respectively assigned to $p_t$ and $1 - p_t$. This transformation enables a comparison between our proposed algorithm, MixUCB, and another optimism-based algorithm that can utilize knowledge of the event probability. This becomes particularly significant when the event probabilities vary over time.

Please note that both UCB1 and "MixUCB with probability estimation" are not well-suited for time-varying environments. In future studies, it is recommended to conduct a more comprehensive empirical comparison with algorithms specifically designed for non-stationary environments, such as those proposed by Galichet et al. [2013] or Slivkins and Upfal [2008]. However, it should be acknowledged that existing techniques may not be directly applicable to the conditions presented in the current experiments.

[2] The code for these experiments can be found at `https://github.com/Thomick/bandit-external-risks`

In this empirical study, we examine a specific instance of a bandit problem with external risk. It involves a 4-armed bandit with Gaussian reward distributions with a variance of 1. Notably, for each arm $i$, we have $\mu_i < \lambda_i$, reflecting the concept of "nominal distribution vs hazard distribution." The intervals of the form $[\lambda_i, \mu_i]$ represent the possible mean rewards for arm $i$ corresponding to different event probabilities. These intervals are visualized in Figure 6. The selection of these intervals ensures that each arm is optimal for a specific value of the event probability $p$. The optimality intervals of the arms are determined by the zeros of the suboptimality gap as a function of $p$, which represents the difference between the best and second-best arms. The variations of this suboptimality gap with respect to the event probability are depicted in Figure 6. While the suboptimality gap can exhibit more complex variations in general, it is guaranteed to be piecewise linear.

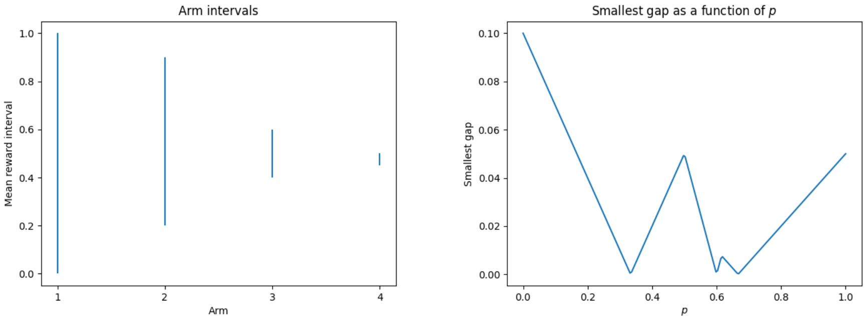


Figure 6: Illustration of the instance of bandit used for experiments. Left: Interval of average reward for each arm when the event probability varies between 0 and 1. Right: Suboptimality gap (average difference between the best and the second-best arm) as a function of the event probability

All experiments were repeated 50 times, and the plots depict the average cumulative rewards, while the error bands represent the standard error of the estimated average cumulative regret. The horizon was set to 10,000, except for the stationary case with $p = 0.4$, where the results appeared inconclusive. Consequently, the horizon was extended. The exact reason for this inconclusiveness is not entirely clear, but it is likely attributed to the narrow suboptimality gap associated with this specific value of the event probability in the given instance.

**Fixed event probability** Figure 7 facilitates the comparison of the four algorithms for a specific value of the event probability, denoted as $p$. We observe that both versions of MixUCB exhibit similar performance to UCB1. However, when the value of $p$ approaches zero, the forced exploration at the beginning of MixUCB becomes apparent and puts the algorithm at a disadvantage compared to UCB1. It seems that longer experimental runs tend to diminish this gap between the two algorithms.

**Periodic event probability** We consider two scenarios: one where the event probability alternates between 0.25 and 0.75, and another where it linearly varies between 0 and 1 over 1000 time steps then repeat. The results of these scenarios are presented in Figure 7. It can be observed that our proposed algorithm, MixUCB, appears to effectively utilize the additional forecast information regarding the event probability. However, it is important to note that a comparison with algorithms specifically designed for time-varying bandits is necessary to draw conclusions about the performance of MixUCB in this non-stationary event probability case.

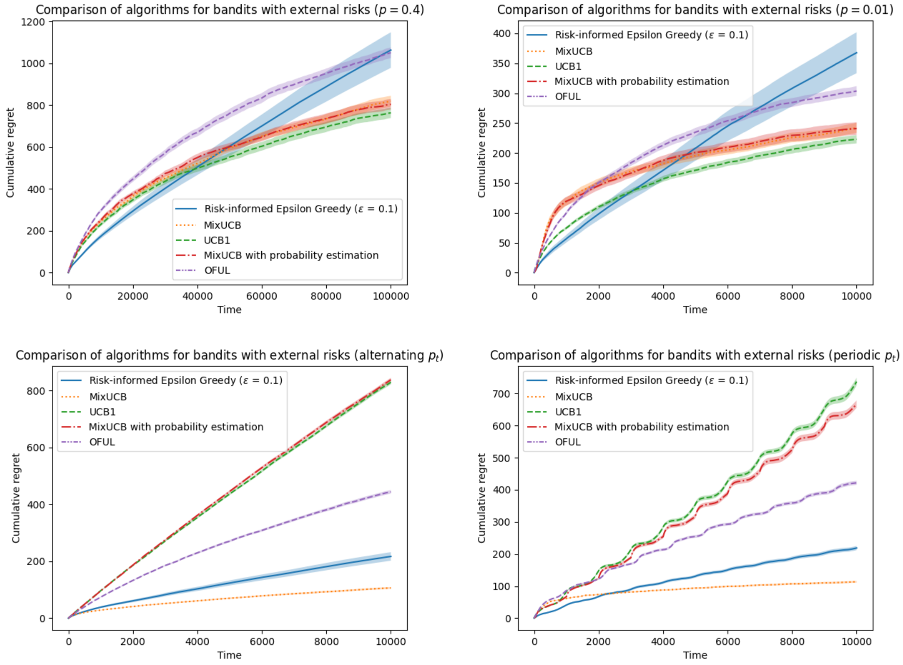


Figure 7: Comparison of four algorithms on an instance of bandits with external risks. The difference between the plots is the considered sequence of event probability $(p_t)_{t=1}^n$. Top row: Fixed event probability. Bottom left: alternance between 0.25 and 0.75. Bottom right: Periodic 1000 steps pattern characterized by a linear increase of event probability between 0 and 1.

**Uniformly sampled event probability** To demonstrate an approach mentioned in the previous section, we conducted a comparison of algorithms within a setting where the event probabilities at each step are uniformly sampled from the range of 0 to 1. The results are presented in Figure 8. The sequences of event probabilities remain consistent across all algorithms, but they differ across consecutive runs. As anticipated, UCB1 and the variant of MixUCB with probability estimation (which both assume stationary reward distributions exhibit linear regret, while MixUCB displays what appears to be sublinear regret. However, it is important to note that a natural criticism is the absence of a meaningful comparison to algorithms designed for non-stationary bandits, which necessitates further investigation in future studies.

**Converging sequence of event probability** We then proceed to empirically investigate the progression of the number of suboptimal pulls when the sequence of event probabilities converges to an optimal arm switching point. For this purpose, we focus on the two-arm bandit example outlined in Section 4.5. The key distinction in our study is that we manipulate the convergence rate of the sequence $(p_t)_{t=1}^n$, resulting in different sequences of gaps $\Delta(t)$. Figure 8 demonstrates the number of

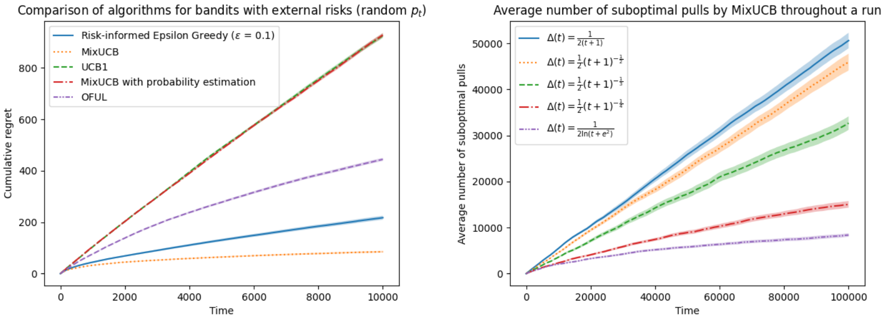


Figure 8: Left: Comparison of algorithms in the case where the event probabilities are uniformly sampled between 0 and 1. Right: Evolution of the number of suboptimal pulls throughout a run depending on the convergence rate toward a switching point.

suboptimal pulls performed throughout the run for various convergence rates. The straight lines in the figure do not directly indicate that the total number of pulls is linear in the horizon, but rather that the arms were sampled at a constant rate throughout the runs. Remarkably, we observe that when the convergence rate surpasses $\frac{1}{\sqrt{t}}$, we start to observe what appears to be a consistent rate of sampling for the suboptimal arm. In the case of slower convergence rates, the results are not as straightforward, but we still observe significantly fewer suboptimal samples.

## 5 Conclusion

The study presents a new mathematical model for forest growth that incorporates interactions between individual trees, as well as multiple environmental hazards with distinct dynamics. The model can function autonomously as a simulated environment for reinforcement learning, adhering to standard programming interfaces. The findings highlight the limitations of implementing a global strategy and underscore the necessity for more sustainable and resilient forest management practices. The study demonstrates the feasibility of uncovering tree-level strategies through reinforcement learning techniques and encourages future investigations into risk-sensitive approaches.

The research emphasizes the significance of adapting strategies based on the actual level of risk and the risk preferences of forest managers. Policies that perform well under specific frequencies of environmental hazards may exhibit decreased performance compared to robust and risk-averse alternatives. A robust forest management policy should demonstrate good performance despite discrepancies between the model and reality.

Future research directions involve exploring variations of the model that incorporate additional complexities, such as multiple species with distinct growth dynamics and interactions. Additionally, incorporating weighted interaction graphs based on proximity and imposing limitations on visible information and manageable areas could provide further insights. These variations address computational challenges and facilitate the study of learning algorithms in this environment.

In addition, we propose a modification to the standard reinforcement learning setting to model the impact of external risks. This modification introduces a framework where random events occasionally and temporarily modify the system's dynamics. The agent receives a forecast estimating the probability of such events, enabling preventive actions to minimize losses. The occurrence of events happens independently of the actual system state, and an effective strategy can be devised by combining the normal dynamics with the dynamics during extreme events. The report focuses on mapping this problem to learning optimal actions within the context of stochastic multi-armed bandits with external risks and provides elements for the study of an algorithm tailored for this setting. We provide an analysis of the algorithm in the case where the probability of extreme events remains constant, and empirically demonstrate promising results in the case in which this probability changes overtime.

Future work on this topic includes a comprehensive analysis of the algorithm in the non-stationary case, which is crucial for solving the problems that motivated this framework in the first place. An extension of the framework to a Markov decision process is a natural progression, allowing events to also change the transition probabilities of the model. This new case may require considering longer-term forecasts, enabling the algorithm to take proactive actions. However, the provision of early forecasts raises the question of forecast uncertainty (forecasts may be less accurate if provided too far in advance). An additional extension of this work would be to consider mixtures of more than two distributions, allowing for the modeling of environments affected by multiple external risks. For example, in a forest environment, one could consider a mixture of distributions representing the risks of forest fires, storms, and diseases. By incorporating multiple risk factors, the decision-making process can become more complex and challenging. This extension would provide a more comprehensive framework for analyzing and optimizing management strategies in environments with multiple overlapping risks.

# References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24, 2011.

Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47:235–256, 2002.

Debabrota Basu, Odalric-Ambrym Maillard, and Timothée Mathieu. Bandits corrupted by nature: Lower bounds on regret and robust optimistic algorithm. *arXiv preprint arXiv:2203.03186*, 2022.

Dorian Baudry, Romain Gautron, Emilie Kaufmann, and Odalric Maillard. Optimal thompson sampling strategies for support-aware cvar bandits. In *International Conference on Machine Learning*, pages 716–726. PMLR, 2021.

Yinlam Chow, Aviv Tamar, Shie Mannor, and Marco Pavone. Risk-sensitive and robust decision-making: a cvar optimization approach. *Advances in neural information processing systems*, 28, 2015.

Stéphane Couture, Marie-Josée Cros, and Régis Sabbadin. Risk aversion and optimal management of an uneven-aged forest under risk of windthrow: A markov decision process approach. *Journal of Forest Economics*, 25:94–114, 2016.

Stéphane Couture, Marie-Josée Cros, and Régis Sabbadin. Multi-objective sequential forest management under risk using a markov decision process-pareto frontier approach. *Environmental Modeling & Assessment*, 26(2):125–141, 2021.

Wesley Cowan, Junya Honda, and Michael N Katehakis. Normal bandits of unknown means and variances. *The Journal of Machine Learning Research*, 18(1):5638–5665, 2017.

Nicolas Galichet, Michele Sebag, and Olivier Teytaud. Exploration vs exploitation vs safety: Risk-aware multi-armed bandits. In *Asian Conference on Machine Learning*, pages 245–260. PMLR, 2013.

Javier Garcıa and Fernando Fernández. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16(1):1437–1480, 2015.

Aurélien Garivier and Eric Moulines. On upper-confidence bound policies for non-stationary bandit problems. *arXiv preprint arXiv:0805.3415*, 2008.

Romain Gautron, Dorian Baudry, Myriam Adam, Gatien N Falconnier, and Marc Corbeels. Towards an efficient and risk aware strategy for guiding farmers in identifying best crop management. *arXiv preprint arXiv:2210.04537*, 2022.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Patrice Loisel. Impact of storm risk on faustmann rotation. *Forest Policy and Economics*, 38: 191–198, 2014.

Patrice Loisel, Marielle Brunette, and Stéphane Couture. Ambiguity, value of information and forest rotation decision under storm risk. Working Papers of BETA 2022-26, Bureau d'Economie Théorique et Appliquée, UDS, Strasbourg, 2022. URL https://ideas.repec.org/p/ulp/sbbeta/2022-26.html.

Odalric-Ambrym Maillard. Robust risk-averse stochastic multi-armed bandits. In *Algorithmic Learning Theory: 24th International Conference, ALT 2013, Singapore, October 6-9, 2013. Proceedings 24*, pages 218–233. Springer, 2013.

Oliver Mihatsch and Ralph Neuneier. Risk-sensitive reinforcement learning. *Machine learning*, 49: 267–290, 2002.

Juan Manuel Morales, Mónica Mermoz, Juan Haridas Gowda, and Thomas Kitzberger. A stochastic fire spread model for north patagonia based on fire occurrence maps. *Ecological Modelling*, 300: 73–80, 2015.

R Tyrrell Rockafellar, Stanislav Uryasev, et al. Optimization of conditional value-at-risk. *Journal of risk*, 2:21–42, 2000.

Nadja Rüger, Uta Berger, Stephen P Hubbell, Ghislain Vieilledent, and Richard Condit. Growth strategies of tropical tree species: disentangling light and size effects. *PloS one*, 6(9):e25330, 2011.

Patrick Saux and Odalric Maillard. Risk-aware linear bandits with convex loss. In *International Conference on Artificial Intelligence and Statistics*, pages 7723–7754. PMLR, 2023.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Aleksandrs Slivkins and Eli Upfal. Adapting to a changing environment: the brownian restless bandits. In *COLT*, pages 343–354, 2008.

Aviv Tamar, Yonatan Glassner, and Shie Mannor. Optimizing the cvar via sampling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

Vincent YF Tan, Krishna Jagannathan, et al. A survey of risk-aware multi-armed bandits. *arXiv preprint arXiv:2205.05843*, 2022.

Iñigo Urteaga and Chris H Wiggins. Nonparametric gaussian mixture models for the multi-armed bandit. *arXiv preprint arXiv:1808.02932*, 2018.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

# Appendices

## A Proof of Theorem 1

**Notations :**

- $\nu_i(p) = (1-p)\mu_i + p\lambda_i$ is the expected reward of arm $i$ given the mixture weight $p$
- $\Delta_i(p) = \max_j \nu_j(p) - \nu_i(p)$ is the sub-optimality gap of arm $i$ given the mixture weight $p$
- $\hat{\nu}_i(t,p)$ is the estimated mean of arm $i$ at time $t$ for the mixture parameter $p$
- $\hat{\nu}_{iu_i}(p)$ is the estimated mean of arm $i$ after $u_i$ pulls for the mixture parameter $p$
- $T_i(n)$ is the number of pull on arm $i$ during the $n$ first rounds
- $T^{\mu}_{iu_i}$ (resp. $T^{\mu}_{iu_i}$) is the number of samples from the distribution with mean $\mu$ (resp. $\lambda$) after $u_i$ pulls of arm $i$

*Proof.* The analysis is adapted from the proof of Theorem 7.1 of Lattimore and Szepesvári [2020]. Without loss of generality, we assume that the first arm is optimal with respect to $p$, meaning that $\nu_1(p) = \max_j \nu_j(p)$. In addition, we will assume that $0 < p < 1$. In the other case, as explained during the definition of the algorithm, the term associated to the impossible event is removed, so the algorithm will be equivalent to UCB($\delta$).

As in the standard stochastic bandit setting, the regret can be decomposed as follows,

$$R_n = \sum_{i:\Delta_i(p)>0} \Delta_i(p)\mathbb{E}[T_i(n)] \tag{7}$$

We establish the result by bounding $T_i(n)$ for each suboptimal arm $i$. Throughout the following analysis, we assume that we have sampled from each distribution of each arm at least once. The average number of samples required from each arm for this event to occur is less than $\frac{1}{p} + \frac{1}{1-p}$, which will be taken into account at the end of the proof.

Let us define a "good" event $G_i$.

$$G_i = \left\{ \nu_1(p) < \min_{t\in[n]} \mathrm{UCB}_1(t,\delta,p) \right\} \cap \left\{ \hat{\nu}_{iu_i}(p) + \sqrt{2\log\left(\frac{1}{\delta}\right)\left(\frac{(1-p)^2}{T^{\mu}_{iu_i}} + \frac{p^2}{T^{\lambda}_{iu_i}}\right)} < \nu_1(p) \right\}$$

where $u_i$ is a constant to be chosen later

We will show that if $G_i$ occurs then $T_i(n)$ is bounded by $u_i$ and the complementary event $G_i^c$ happens only with low probability. This will allow us to bound $\mathbb{E}[T_i(n)]$,

$$\mathbb{E}\left[T_i(n)\right] = \mathbb{E}\left[\mathbb{I}\left\{G_i\right\} T_i(n)\right] + \mathbb{E}\left[\mathbb{I}\left\{G_i^c\right\} T_i(n)\right] \leqslant u_i + \mathbb{P}\left(G_i^c\right) n \tag{8}$$

Assume that $G_i$ holds, let us show by contradiction that $T_i(n) \leqslant u_i$. First suppose that $T_i(n) > u_i$, if that the case then that means there exists a round $t \leqslant n$ where $T_i(t-1) = u_i$ and the selected action is $i$. By definition of $G_i$ we have the following inequality,

$$\begin{aligned}
\mathrm{UCB}_i(t-1,\delta,p) &= \hat{\nu}_i(t-1,p) + \sqrt{2\log\left(\frac{1}{\delta}\right)\left(\frac{(1-p)^2}{T_i^{\mu}(t-1)} + \frac{p^2}{T_i^{\lambda}(t-1)}\right)} \\
&= \hat{\nu}_{iu_i}(p) + \sqrt{2\log\left(\frac{1}{\delta}\right)\left(\frac{(1-p)^2}{T_{iu_i}^{\mu}} + \frac{p^2}{T_{iu_i}^{\lambda}}\right)} && (\text{ since } T_i(t-1) = u_i) \\
&< \nu_1(p) && (\text{ definition of } G_i) \\
&< \mathrm{UCB}_1(t-1,\delta,p). && (\text{ definition of } G_i)
\end{aligned}$$

However the above inequality contradict the fact that the algorithm selected the arm $i$ since there exists an arm with a higher index. Therefore, if $G_i$ holds then $T_i(n) \leqslant u_i$.

Hence, there remains to bound the probability of each part of the complementary event $G_i^c$

$$G_i^c = \left\{\nu_1(p) \geqslant \min_{t\in[n]} \mathrm{UCB}_1(t,\delta,p)\right\} \cup \left\{\hat{\nu}_{iu_i}(p) + \sqrt{2\log\left(\frac{1}{\delta}\right)\left(\frac{(1-p)^2}{T_{iu_i}^{\mu}} + \frac{p^2}{T_{iu_i}^{\lambda}}\right)} \geqslant \nu_1(p)\right\} \tag{9}$$

The probability of the first part can be upper bounded using the definition of $\mathrm{UCB}_1(t,\delta,p)$ and a union bound. The last inequality follows from Proposition 2.

$$\begin{aligned}
\mathbb{P}\left(\nu_1(p) \geqslant \min_{t\in[n]} \mathrm{UCB}_1(t,\delta,p)\right) &\leqslant \mathbb{P}\left(\nu_1(p) \geqslant \min_{s,s'\in[n]} \hat{\nu}_{i(s+s')}(p) + \sqrt{2\log\left(\frac{1}{\delta}\right)\left(\frac{(1-p)^2}{s} + \frac{p^2}{s'}\right)}\right) \\
&\leqslant \mathbb{P}\left(\bigcup_{s,s'\in[n]} \left\{\nu_1(p) \geqslant \hat{\nu}_{i(s+s')}(p) + \sqrt{2\log\left(\frac{1}{\delta}\right)\left(\frac{(1-p)^2}{s} + \frac{p^2}{s'}\right)}\right\}\right) \\
&\leqslant \sum_{s,s'\in[n]} \mathbb{P}\left(\left\{\nu_1(p) \geqslant \hat{\nu}_{i(s+s')}(p) + \sqrt{2\log\left(\frac{1}{\delta}\right)\left(\frac{(1-p)^2}{s} + \frac{p^2}{s'}\right)}\right\}\right) \\
&\leqslant \sum_{s,s'\in[n]} \delta \leqslant n^2\delta
\end{aligned} \tag{10}$$

As for the second part of $G_i^c$, we use the formula of total probability to decouple the randomness introduced by the random choice between the two distributions associated to each arm from the one of the index computation. We then cut the sum into three parts. The two ends are bounded by the tail probability of a binomial distribution and the middle term is bound separately.

$$
\begin{aligned}
&\mathbb{P}\left(\hat{\nu}_{iu_i}(p)+\sqrt{2\log\left(\frac{1}{\delta}\right)\left(\frac{(1-p)^2}{T^{\mu}_{iu_i}}+\frac{p^2}{T^{\lambda}_{iu_i}}\right)}\geqslant \nu_1(p)\right)\\
&\leqslant \sum_{s=0}^{u_i}\mathbb{P}\left(\hat{\nu}_{iu_i}(p)+\sqrt{2\log\left(\frac{1}{\delta}\right)\left(\frac{(1-p)^2}{s}+\frac{p^2}{u_i-s}\right)}\geqslant \nu_1(p)\right)\mathbb{P}\left(T^{\mu}_{iu_i}=s\right) \qquad \text{(Total probability)}\\
&\leqslant \sum_{s=0}^{u_i}\mathbb{P}\left(\hat{\nu}_{iu_i}(p)-\nu_i(p)+\geqslant \Delta_i(p)-\sqrt{2\log\left(\frac{1}{\delta}\right)\left(\frac{(1-p)^2}{s}+\frac{p^2}{u_i-s}\right)}\right)\mathbb{P}\left(T^{\mu}_{iu_i}=s\right)\\
&\leqslant \sum_{s=0}^{v_i^l}\mathbb{P}\left(T^{\mu}_{iu_i}=s\right)+\sum_{s=v_i^l+1}^{u_i-v_i^r-1}\mathbb{P}\left(\hat{\nu}_{iu_i}(p)-\nu_i(p)+\geqslant \Delta_i(p)-\sqrt{2\log\left(\frac{1}{\delta}\right)\left(\frac{(1-p)^2}{s}+\frac{p^2}{u_i-s}\right)}\right)\\
&\quad+\sum_{s=u_i-v_i^r}^{u_i}\mathbb{P}\left(T^{\mu}_{iu_i}=s\right) \qquad (11)
\end{aligned}
$$

where $v_i^l$ and $v_i^r$ are chosen such that for a value $u_i$ sufficiently large and for all $s$ such that $v_i^l \leqslant s \leqslant u_i - v_i^r$ we have

$$
\Delta_i(p)-\sqrt{2\log\left(\frac{1}{\delta}\right)\left(\frac{(1-p)^2}{s}+\frac{p^2}{u_i-s}\right)}\geqslant c\Delta_i(p) \qquad (12)
$$

where $c$ is a constant to be chosen later.

For such a choice of $v_i^l$ and $v_i^r$ and by applying a Chernoff bound to each of the term in the middle sum of Eq. 11 we obtain

$$
\begin{aligned}
&\sum_{s=v_i^l+1}^{u_i-v_i^r-1}\mathbb{P}\left(\hat{\nu}_{iu_i}(p)-\nu_i(p)\geqslant \Delta_i(p)-\sqrt{2\log\left(\frac{1}{\delta}\right)\left(\frac{(1-p)^2}{s}+\frac{p^2}{u_i-s}\right)}\right)\\
&\leqslant \sum_{s=v_i^l+1}^{u_i-v_i^r-1}\mathbb{P}\left(\hat{\nu}_{iu_i}(p)-\nu_i(p)\geqslant c\Delta_i(p)\right)\\
&\leqslant \sum_{s=v_i^l+1}^{u_i-v_i^r-1}\exp\left(\frac{-c^2\Delta_i(p)^2}{2\left(\frac{(1-p)^2}{s}+\frac{p^2}{u_i-s}\right)}\right)\\
&=(u_i-v_i^l-v_i^r-2)\delta^{\frac{c^2}{(1-c)^2}} \qquad \leqslant u_i\delta^{\frac{c^2}{(1-c)^2}} \qquad (13)
\end{aligned}
$$

We identify the two other sums of Eq. 11 as tail probabilities of Binomial distributions and we bound them by applying Hoeffding inequality.

$$\sum_{s=0}^{v_i^l} \mathbb{P}\left(T^{\mu}_{iu_i} = s\right) \leqslant \exp\left(-2u_i\left(1-p-\frac{v_i^l}{u_i}\right)^2\right)$$
$$\sum_{s=u_i-v_i^r}^{u_i} \mathbb{P}\left(T^{\mu}_{iu_i} = s\right) \leqslant \exp\left(-2u_i\left(p-\frac{v_i^r}{u_i}\right)^2\right)$$

It is worth noting that the above sums become null when $v_i^l$ (respectively $v_i^r$) equals zero. This is because we initially assumed that each arm's distribution was sampled at least once, resulting in the remaining terms in the sum being zero probabilities.

Now we can choose the value of the constants left behind,

$$u_i = \left\lceil \frac{4\log(1/\delta)}{(1-c)^2\Delta_i^2} \right\rceil$$
$$v_i^l = \left\lfloor \frac{(1-p)u_i}{2} \right\rfloor$$
$$v_i^r = \left\lfloor \frac{pu_i}{2} \right\rfloor$$

We can verify with this particular choice, inequality 12 holds. Now by putting everything together we obtain

$$T_i(p) \leqslant \frac{4\log(1/\delta)}{(1-c)^2\Delta_i^2} + n\left(n^2\delta + \frac{u_i}{2}\delta^{\frac{c^2}{(1-c)^2}} + \delta^{\frac{2(1-p)^2}{(1-c)^2\Delta_i^2}} + \delta^{\frac{2p^2}{(1-c)^2\Delta_i^2}}\right)$$

Choosing arbitrarily $c = 1/2$ and replacing $\delta = \frac{1}{n^3}$ yield the desired result.

□

# B Proof of Theorem 2

*Proof.* Without loss of generality, we assume that the first arm is optimal with respect to $p$, meaning that $\nu_1(p) = \max_j \nu_j(p)$.

We prove the result by bounding $T_i(n)$ for each suboptimal arm $i$. Let us define $G_i(q)$ an event indicating the correct estimation of the parameters after a certain number of pulls, given a mixture parameter $q$.

$$G_i(q) = \left\{\nu_1(q) < \min_{t\in[n]} \mathrm{UCB}_1(t,\delta,q)\right\} \cap \left\{\hat{\nu}_{iu_i}(q) + \sqrt{2\log\left(\frac{1}{\delta}\right)\left(\frac{(1-q)^2}{T^{\mu}_{iu_i}} + \frac{q^2}{T^{\lambda}_{iu_i}}\right)} < \nu_1(q)\right\}$$

where $u_i$ is a constant to be chosen later

We start by decomposing $T_i(n)$ using indicator functions.

$$
\begin{aligned}
T_i(n) = & \sum_{t=1}^{n} \mathbb{1}\{A_t = i\}\mathbb{1}\{G_i(\hat{p}_t)\} && (14)\\
& + \sum_{t=1}^{n} \mathbb{1}\{A_t = i\}\mathbb{1}\{G_i^c(\hat{p}_t)\}\mathbb{1}\{|\hat{p}_t - p| \leqslant \varepsilon_t\} && (15)\\
& + \sum_{t=1}^{n} \mathbb{1}\{A_t = i\}\mathbb{1}\{G_i^c(\hat{p}_t)\}\mathbb{1}\{|\hat{p}_t - p| > \varepsilon_t\} && (16)
\end{aligned}
$$

By linearity, we can now bound the expected value of $T_i(n)$ by the sum of the individual expectation of 14, 15 and 16.

We can first notice that 14 is less than $u_i$. Indeed, after pulling arm $i$ $u_i$ times, $\{A_t = i\}$ and $G_i(p_t)$ become incompatible. So $\mathbb{E}(\sum_{t=1}^{n} \mathbb{1}\{A_t = i\}\mathbb{1}\{G_i(\hat{p}_t)\}) \leqslant u_i$.

Now bounding the expectation of 16, we have :

$$
\begin{aligned}
\mathbb{E}(\sum_{t=1}^{n} \mathbb{1}\{A_t = i\}\mathbb{1}\{G_i^c(\hat{p}_t)\}\mathbb{1}\{|\hat{p}_t - p| > \varepsilon_t\}) & \leqslant \sum_{t=1}^{n} \mathbb{P}(|\hat{p}_t - p| > \varepsilon_t)\\
& \leqslant \sum_{t=1}^{n} \exp(-2t\varepsilon_t) && (17)
\end{aligned}
$$

Finally, it remains to bound 15,

$$
\begin{aligned}
&\mathbb{E}(\sum_{t=1}^{n} \mathbb{1}\{A_t = i\}\mathbb{1}\{G_i^c(\hat{p}_t)\}\mathbb{1}\{|\hat{p}_t - p| \leqslant \varepsilon_t)\\
&\qquad \leqslant \sum_{t=1}^{n} \mathbb{P}(G_i^c(\hat{p}_t) \wedge |\hat{p}_t - p| \leqslant \varepsilon_t)\\
&\qquad \leqslant w_i + \sum_{t=w_i}^{n} \mathbb{P}(G_i^c(\hat{p}_t) \wedge |\hat{p}_t - p| \leqslant \varepsilon_t)\\
&\qquad \leqslant w_i + \sum_{t=w_i}^{n} \mathbb{P}(\nu_1(p_t) < \min_{t\in[n]} \mathrm{UCB}_1(t, \delta, p_t) \wedge |\hat{p}_t - p| \leqslant \varepsilon_t)\\
&\qquad\quad + \sum_{t=w_i}^{n} \mathbb{P}\left(\hat{\nu}_{iu_i}(p_t) + \sqrt{2\log\left(\frac{1}{\delta}\right)\left(\frac{(1-p_t)^2}{T_{iu_i}^{\mu}} + \frac{p_t^2}{T_{iu_i}^{\lambda}}\right)} < \nu_1(p_t) \wedge |\hat{p}_t - p| \leqslant \varepsilon_t\right) && (18)
\end{aligned}
$$

where $w_i$ is chosen such that $\forall t \geqslant w_i, \forall q \in [p - \varepsilon_t, p + \varepsilon_t]$, both the following condition are verified,

$$
\Delta_i(q) \geqslant \frac{7\Delta_i(p)}{8} \qquad (19)
$$

and

$$\forall s \in [\![1, u_i]\!]\,, \left|\sqrt{2\log\left(\frac{1}{\delta}\right)\left(\frac{(1-q)^2}{s}+\frac{q^2}{u_i-s}\right)}-\sqrt{2\log\left(\frac{1}{\delta}\right)\left(\frac{(1-p)^2}{s}+\frac{p^2}{u_i-s}\right)}\right| \leqslant \frac{\Delta_i(p)}{8} \tag{20}$$

We then proceed as in the risk-informed case by defining $v_i^l$ and $v_i^r$ such that for a value $u_i$ sufficiently large and for all $s$ such that $v_i^l \leqslant s \leqslant u_i - v_i^r$ we have

$$\Delta_i(p) - \sqrt{2\log\left(\frac{1}{\delta}\right)\left(\frac{(1-p)^2}{s}+\frac{p^2}{u_i-s}\right)} \geqslant c\Delta_i(p) \tag{21}$$

where $c$ is a constant to be chosen later. By definition of $w_i$, the condition 21 also implies that for any $t \in [\![w_i, n]\!]$, if $|\hat{p}_t - p| \leqslant \varepsilon_t$ holds, we have

$$\begin{aligned}\Delta_i(p_t) - \sqrt{2\log\left(\frac{1}{\delta}\right)\left(\frac{(1-p_t)^2}{s}+\frac{p_t^2}{u_i-s}\right)} &\geqslant \frac{\Delta_i(p)}{8} - \sqrt{2\log\left(\frac{1}{\delta}\right)\left(\frac{(1-p)^2}{s}+\frac{p^2}{u_i-s}\right)} - \frac{\Delta_i(p)}{8}\\ &\geqslant (c-\frac{1}{4})\Delta_i(p)\end{aligned}$$

We perform the same decomposition in three part as in the proof of 1. The two extreme terms are identical (they only depend on the chosen $v_i^l$ and $v_i^r$) and we will now handle the middle part for any given $t \geqslant w_i$.

$$\begin{aligned}&\sum_{s=v_i^l+1}^{u_i-v_i^r-1}\mathbb{P}\left(\hat{\nu}_{iu_i}(p_t)+\sqrt{2\log\left(\frac{1}{\delta}\right)\left(\frac{(1-p_t)^2}{T_{iu_i}^{\mu}}+\frac{p_t^2}{T_{iu_i}^{\lambda}}\right)} < \nu_1(p_t) \wedge \hat{p}_t - p| \leqslant \varepsilon_t\right)\\ &\leqslant \sum_{s=v_i^l+1}^{u_i-v_i^r-1}\mathbb{P}\left(\hat{\nu}_{iu_i}(p_t)-\nu_i(p_t) \geqslant \Delta_i(p_t) - \sqrt{2\log\left(\frac{1}{\delta}\right)\left(\frac{(1-p_t)^2}{s}+\frac{p_t^2}{u_i-s}\right)}\right)\\ &\leqslant \sum_{s=v_i^l+1}^{u_i-v_i^r-1}\mathbb{P}\left(\hat{\nu}_{iu_i}(p_t)-\nu_i(p_t) \geqslant (c-\frac{1}{4})\Delta_i(p)\right)\\ &\leqslant \sum_{s=v_i^l+1}^{u_i-v_i^r-1}\exp\left(\frac{-(c-\frac{1}{4})^2\Delta_i(p)^2}{2\left(\frac{(1-p_t)^2}{s}+\frac{p_t^2}{u_i-s}\right)}\right)\\ &= (u_i - v_i^l - v_i^r - 2)\delta^{\frac{(c-\frac{1}{4})^2}{(1-c)^2}}\\ &\leqslant u_i\delta^{\frac{(c-\frac{1}{4})^2}{(1-c)^2}}\end{aligned} \tag{22}$$

We now set $\varepsilon_t = \sqrt{\frac{\log(n)}{t}}$ such that 17 is bounded by $\frac{1}{n}$.

The constants are chosen as in the proof of Theorem 1 :

$$u_i = \left\lceil \frac{4\log(1/\delta)}{(1-c)^2\Delta_i^2} \right\rceil$$
$$v_i^l = \left\lfloor \frac{(1-p)u_i}{2} \right\rfloor$$
$$v_i^r = \left\lfloor \frac{pu_i}{2} \right\rfloor$$

It can then be shown that condition 19 is verified when

$$t \geqslant \frac{16|\mu_i - \lambda_i + \lambda - \mu_1|^2 \log(n)}{9\Delta_i^2}$$

while condition 20 holds when

$$t \geqslant \frac{64\log(n)\log^2(1/\delta)}{(1-c)^2}$$

We chose $w_i$ to be the maximum between these two values.

Finally, by assembling the bounds from 17,18 and 22 and replacing the constants we obtain the announced result of Theorem 2 □