# Two-Armed Restless Bandits with Imperfect Information: Stochastic Control and Indexability*

Roland G. Fryer, Jr.
Harvard University and NBER

Philipp Harms
ETH Zurich

May 2015

**Abstract**

We present a two-armed bandit model of decision making under uncertainty where the expected return to investing in the "risky arm" increases when choosing that arm and decreases when choosing the "safe" arm. These dynamics are natural in applications such as human capital development, job search, and occupational choice. Using new insights from stochastic control, along with a monotonicity condition on the payoff dynamics, we show that optimal strategies in our model are stopping rules that can be characterized by an index which formally coincides with Gittins' index. Our result implies the indexability of a new class of restless bandit models.

*We are grateful to Richard Holden, Peter Michor, Derek Neal, Ariel Pakes, Yuliy Sannikov, Mete Soner, Josef Teichmann and seminar participants at Barcelona GSE and Harvard University for helpful comments. Financial support from the Education Innovation Laboratory at Harvard University is gratefully acknowledged. Correspondence can be addressed to the authors by e-mail: rfryer@fas.harvard.edu [Fryer] or pharms@math.ethz.ch [Harms]. The usual caveat applies.

# 1 Introduction.

Bandit models are decision problems where, at each instant of time, a resource like time, effort, or money has to be allocated strategically between several options, referred to as the arms of the bandit. When selected, the arms yield payoffs that typically depend on unknown parameters. Arms that are not selected remain unchanged and yield no payoff. The key idea in this class of models is that agents face a tradeoff between experimentation (gathering information on the returns to each arm) and exploitation (choosing the arm with the highest expected value).

Over the past sixty years, bandit models have become an important framework in economic theory, applied mathematics and probability, and operations research. They have been used to analyze problems as diverse as market pricing, the optimal design of clinical trials, product search and the research and development activities of firms (Rothschild [56], Berry and Fristedt [7], Bolton and Harris [8], and Keller and Rady [30]). To understand how firms set prices without a clear understanding of their demand curves, Rothschild [56] posits that firms repeatedly charge prices and observe the resulting demand. Setting prices too high or too low is costly for firms (experimentation), but allows them to learn about the optimal price (exploitation). In the optimal design of clinical trials, Berry and Fristedt [7] formulate the problem as: given a fixed research budget, how does one allocate effort among competing projects, whose properties are only partially known at a given point in time but may be better understood as time passes. In product search, customers sample products to learn about their quality. Their optimizing behavior can be described as in Bolton and Harris [8, 9]. In these models, news about the quality of the product arrive continuously. The situation where news arrive only occasionally, e.g. in the form of break-throughs in research, is modeled by Keller et al. [31, 30].

An important assumption in the classical bandit literature is that the reward distribution of arms that are not chosen does not evolve; they rest (Gittins, Glazebrook, and Weber [21]). This assumption seems natural in many applications. Yet, in many other important scenarios, it seems overly restrictive.[1] Consider, for instance, the possibility of dynamic complementarities in human capital production.[2] Imagine a student who has the choice of whether or not to invest effort into her school work. Today's effort is rewarded by being more at ease with tomorrow's course work, or the ability to glean a deeper understanding from class lectures. As Cunha and Heckman [11] note, "learning begets learning." Conversely, not doing one's assignments today might give instantaneous gratification, but makes tomorrow's school work harder. More generally, this dynamic can be found in the context of human capital formation when early investments in human capital increase the expected payoff of future investments, while a lack of early investments has the reverse effect. These dynamics require arms that evolve, even when they are not used.

As a second example, consider an unemployed worker looking for a job. With every job application, she gathers both information about the job market and experience in the application process,

[1] The importance of relaxing this assumption has been recognized early on in the seminal work of Whittle [68], who proposed clinical trials, aircraft surveillance, and assignment of workers to tasks as potential applications.

[2] Cunha et al. [12] make a similar argument in a different context.

which typically increases her chances of successful future job applications. Conversely, not actively searching for a job may decrease the probability of finding a job in future applications. This could be due to market penalties for unemployment spells, being disconnected from the changing characteristics of the job market and the application process, or be considered a signal of low motivation by potential employers.

Bandits whose inactive arms are allowed to evolve are known as restless bandits.[3] Generally, optimal strategies for restless bandits are unknown.[4] Nevertheless, when a certain indexability condition is met, Whittle's index [68] can lead to approximately optimal solutions (Weber and Weiss [65, 64]). This index plays the same fundamental role for restless bandits that Gittins' index [20] has for classical ones: it decomposes the task of solving multi-armed bandits into multiple tasks of solving bandits with one safe and one risky arm. The safe arm yields constant rewards and can be interpreted as a cost of investment in the risky arm. Deriving conditions that identify general classes of indexable restless bandit models is an important contribution – permitting more complete analysis of decision problems in which choices jointly effect instantaneous payoffs as well as the distribution of those payoffs in the future – and the subject of this paper.

The origins of this work are the classical bandit models of Bolton and Harris [8], Keller and Rady [30], and Cohen and Solan [10], that we extend to the restless case. In these works, the reward from the risky arm is Brownian motion, a Poisson process, or a Levy process. The unobserved quantity is a Bernoulli variable. Our model is an extension of these models containing them as special cases.[5] Namely, we allow the same generality of reward processes with both volatility and jumps, but make the reward distribution dependent on the type of the agent and the history of past investments. The latter dependence is mediated by a real valued variable that increases while the agent invests in the risky arm and decreases otherwise. In line with our motivating examples of human capital formation and job search, we call this variable the agent's human capital.

The bandit model is first formulated as a problem of stochastic optimal control under *partial observations* in continuous time.[6] Standard formulations of the control problem with partial observations do not work for restless bandit models. However, we show that the frameworks of Fleming and Nisio [18], Wonham [69], and Kohlmann [33] can be used and extended to general controlled Markov processes. We describe these issues in detail in Section 2.2, since they are rarely discussed in the context of bandit problems.

The first result in this paper is a *separation theorem* (Theorem 1) that establishes the equivalence of the control problem with partial observations to a control problem with full observations called the separated control problem. This equivalence is crucial for the solution of the problem and is

[3]Bandits where the active and passive action have opposite effects on payoffs are called *bi-directional bandits* (Glazebrook, Kirkbride, and Ruiz-Hernandez [22]), and our model falls into this class.

[4]Numerical solutions can be obtained by (possibly approximate) dynamic programming or a linear programming reformulation of the problem (Kushner and Dupuis [38], Powell [52], and Nino-Mora [46]).

[5]However, some of these works focus on strategic equilibria involving multiple agents, whereas we only treat the single agent case.

[6]Modeling time as continuous allows one to treat discrete-time models with varying step sizes in a unified framework. We show in Theorem 1 that discrete-time versions of the model converge to the continuous-time limit. This is not true in some other and recent approaches (see Remarks 1 and 3).

implicitly used in many works, including Bolton and Harris [8], Keller and Rady [30], and Cohen and Solan [10]. The separated problem is derived from the partially observable one by replacing the unobserved quantity by its filter, which is its conditional distribution given the past observations. Put differently, the filter is the belief of the agent in being of the high type. In the separated problem, admissibility of controls is defined without the strong measurability constraints present in the control problem with partial observations. Therefore, standard results about the existence of optimal controls and the equivalence to dynamic and linear programming can be applied.

Our second, and main, result (Theorem 2) is the *optimality of stopping rules*, meaning that it is always better to invest first in the risky arm and then in the safe arm instead of the other way round. This result hinges on the monotonic dependence of payoffs on past investment. Intuitively, the sequence of investments matters for two reasons. First, investments in the risky arm reveal information about the distribution of future rewards. The sooner this information becomes available, the better. Second, early investments in the safe arm deteriorate the rewards of later investments in the risky arm. By contrast, early investments in the risky arm do not make the safe arm any less profitable.

We present an unconventional approach to show the optimality of stopping rules. The work horse of most of the bandit literature is either the Hamilton-Jacobi-Bellman (HJB) equation or a setup using time changes. The inclusion of human capital as a state variable turns the HJB equation into a second order partial differential-difference equation. It seems unlikely that explicit solutions of this equation can be found. Moreover, the approach using time changes is not well adapted to the new dynamics of our model. We circumvent these difficulties by investigating the sample paths of optimal strategies. More specifically, we discretize the problem in time and show that any optimal strategy can be modified such that the agent never invests after a period of not investing and such that the modified strategy is still optimal. This interchange argument has been originally developed by Berry and Fristedt [7] for classical bandits. It turns out that the monotonic dependence of the payoffs on the amount of past investment is exactly what is needed to generalize the argument to restless bandits.

Once the optimality of stopping rules is established, it follows easily that optimal strategies can be characterized by an *index* rule. Formally, the index is the same as the one proposed in the celebrated result by Gittins [20] on classical bandits, but inactive arms are allowed to evolve. The explicit formula for the index yields comparative statics of the frontier with respect to the parameters of the model. Most importantly, subsidies of the safe arm enlarge the set of states where the safe arm is optimal, which means that our bandit model is *indexable* in the sense of Whittle [68]. More generally, any arm of a multi-armed restless bandit that satisfies our monotonicity condition is indexable. To our knowledge, this is the first time that a sufficient condition for indexability of a general class of restless bandits with continuous state space and a corresponding rich class of reward processes has been formulated.[7]

[7]Some sensor management models are indexable and have a continuous state space after their transformation to fully observed Markov decision problems (Washburn [63]). This is, however, not the case in their formulation as control problems with partial observations.

To explain the structure of optimal strategies, we consider how information is processed by agents in our model. We work in a Bayesian setting where the agent has a prior about being either "high" or "low type." Rewards obtained from the risky arm depend on this type and are used by the agent to form a posterior belief. The current levels of belief and human capital determine at each stage whether it is optimal to invest in the risky or safe arm. Namely, there is a curve in the belief–human capital domain such that it is optimal to invest in the risky arm if the current level of belief and human capital lies to the right and above the curve. Otherwise, it is optimal to invest in the safe arm. The curve is called the *decision frontier*.

Similar to classical bandit models, the region below the decision frontier is absorbing: agents do not obtain any new information, and their belief remains constant, while their human capital decreases continually. There is, however, an important, and potentially empirically relevant, difference. Not only is the safe arm absorbing – it is depreciating; agents drift further and further away from the frontier. Empirically, this implies that there are very few "marginal" agents. Programs (e.g. lower class size, school choice, financial incentives) designed to increase student achievement at the margin are likely to be ineffective unless: (a) they are initiated when students get close to the decision frontier, or (b) force inframarginal students to invest in the risky arm (e.g. some charter schools, see Dobbie and Fryer [14]). Consistent with Cunha et al. [12], our model predicts that, on average, the longer society waits to invest, the more aggressive the investment needs to be.

The situation is different for agents above the frontier. They continually obtain new information about their type and update their posterior belief accordingly. At the same time, their level of human capital increases. In the long run, there are two possibilities. Either there is some point in time where they hit the frontier. This happens when they encounter a series of bad outcomes from the risky arm and their belief level drops down far enough. In this case, they meet the same fate as agents who originally started out below the frontier. Or they never reach the frontier. In this case, they invest in the risky arm forever and learn their true type in the limit. In fact, under reasonable assumptions, investing in the risky arm for an infinite amount of time is necessary and sufficient for *asymptotic learning* to occur (see Proposition 3). To summarize, agents of the low type eventually end up choosing the safe arm, which is optimal for them. High type agents, however, can get discouraged by bad luck and stop investing in the risky arm, even though the risky arm would be optimal.

The paper is structured as follows. Section 2 provides a brief review of the bandit literature in economics and applied mathematics. Section 3 contains the definitions of the control problems and the separation theorem. Section 3 specializes the general framework of the previous section to restless bandit models satisfying the monotonicity condition and establishes the optimality of stopping rules characterized by Gittins' index. Finally, Section 5 concludes.

# 2 Previous literature.

## 2.1 Bandit models.

Originally developed by Robbins [55], bandit models have been used to analyze a wide range of economic and applied math problems.[8] The first paper where a bandit model was used in an economic context is Rothschild [56], in which a single firm facing a market with unknown demand has to determine optimal prices. Subsequent applications of bandit models include partner search, effort allocation in research, clinical trials, network scheduling and voting in repeated elections (McCall and McCall [44], Weitzman [66], Berry and Fristedt [7], Li and Neely [41], and Banks and Sundaram [3]).

Classical bandits with reward processes driven by Brownian motion or a Poisson process were first solved by Karatzas [28] and Presman [54]. Subsequently, Bolton and Harris [8, 9] and Keller e.a. [31, 30, 29] derived explicit formulas for optimal strategies in the case where the unobservable quantity is a Bernoulli variable and treated strategic interactions of multiple agents. Cohen and Solan [10] unified the formulas obtained for the single agent case and solved a bandit model where the reward is driven by a Levy process with unknown Levy triplet.

Many extensions and variations of classical bandit problems have been proposed, including: bandits with a varying finite or infinite numbers of arms (Whittle [67] and Banks and Sundaram [3]), bandits where an adversary has control over the payoffs (Auer et al. [2]), bandits with dependent arms (Pandey, Chakrabarti, and Agarwal [48]), bandits where multiple arms can be chosen at the same time (Whittle [68]), bandits whose arms yield rewards even when they are inactive (Glazebrook, Kirkbride, and Ruiz-Hernandez [22]), and bandits with switching costs (Banks and Sundaram [4]).

One of the most mathematically challenging extensions is to allow inactive arms to evolve. Such bandits are often referred to as "restless bandits."[9] This term was coined in the seminal paper of Whittle [68]. Beyond mathematical intrigue, there are many practical applications: aircraft surveillance, sensor scheduling, queue management, clinical trials, assignment of workers to tasks, robotics, and target tracking (Ny, Dahleh, and Feron [47], Veatch and Wein [62], Whittle [68], Faihe and Müller [17], and La Scala and Moran [39]). In aircraft surveillance, Ny, Dahleh, and Feron [47] discuss the problem of surveying ships for possible bilge water dumping. A group of unmanned aerial vehicles can be sent to the sites of the ships. The rewards are associated with the detection of a dumping event. The problem falls into the class of sensor management problems, where a set of sensors has to be assigned to a larger set of channels whose state evolves stochastically. In queue management, Veatch and Wein [62] consider the task of scheduling a make-to-stock production facility with multiple products. Finished products are stored in an inventory. Too small an inventory risks incurring backorder or lost sales costs, while too large an inventory

[8] Basu, Bose, and Ghosh [5], Bergemann and Välimäki [6], and Mahajan and Teneketzis [42] provide excellent surveys of the literature on bandit models. The monographs by Presman and Sonin [53], Berry and Fristedt [7] and Gittins, Glazebrook, and Weber [21] contain more detailed presentations.

[9] Some bandits with switching costs can be modeled as restless bandits (Jun [27]).

increases holding costs. In robotics, Faihe and Müller [17] consider the behaviors coordination problem in a setting of reinforcement learning: a robot is trained to perform complex actions that are synthesized from elementary ones by giving it feedback about its success.

### 2.2 Optimal control with partial observations.

In control problems with partial observations, strategies are not allowed to depend on the hidden state. To enforce this constraint, one requires them to be measurable with respect to the sigma algebra generated by the observations. In continuous time, this measurability condition is not strong enough to exclude pathological cases like Example 1 in this paper.

This problem was solved in a setting with additive, diffusive noise by requiring the existence of a change of measure, called Zakai's transform (Fleming and Pardoux [19]), which transforms the observation process into standard Brownian motion. Unfortunately, this approach is not amenable to bandit models, where such a change of measure does not exist because the volatility of the observation process depends on the strategy. Another approach, which was applied successfully to classical bandit models, is to define strategies as time changes (El Karoui and Karatzas [15]). Unfortunately, this technique does not work for restless bandit problems, where inactive arms are allowed to evolve.

Our approach can be seen as a generalization of Fleming and Nisio [18], Wonham [69], and Kohlmann [33]. In these works, the strategies are required to be Lipschitz continuous to ensure well-posedness of the corresponding martingale problem. This excludes discontinuous strategies like cut-off rules, which are typically encountered in bandit problems. We replace the Lipschitz condition by the weaker and more direct requirement that the martingale problem is well-posed. The resulting class of admissible strategies is large enough to contain optimal strategies of classical bandit models and of the restless bandit model in Section 4. It is also small enough to exclude degeneracies like Example 1 and to admit approximations in value by piecewise constant controls (see Theorem 1). For piecewise constant controls the definition of admissibility is unproblematic.

### 2.3 Optimality of stopping rules.

For classical bandit models with one safe and one risky arm, the optimality of stopping rules is a well-known result (Berry and Fristedt [7] and El Karoui and Karatzas [15]). Several approaches to establish this result can be found in the literature. In one approach, the rewards of each arm are fixed in advance and strategies are time changes. The reward that is obtained under a strategy is the time change applied to the reward process. This setup, which has been proposed by Mandelbaum [43], allows a very simple formulation of the measurability constraints on the strategies. It is, however, not well-suited to bandits with evolving arms. In a second approach, one solves the Hamilton-Jacobi-Bellman (HJB) equation for the value function. When this succeeds, the explicit form of the value function can be used to establish the optimality of stopping rules (Bolton and Harris [8], Keller, Rady, and Cripps [31], and Cohen and Solan [10]). In our model, however, the dynamics of the reward distribution introduce an additional state variable, which turns the HJB

equation into a non-local partial differential equation which we cannot solve directly. Moreover, the value function might not be a solution in a classical sense. Pham [51, 50] showed that under suitable assumptions, the value function is a viscosity solution of the HJB equation. It remains open how this could be used to show that stopping rules are optimal. The third approach is to rewrite the problem as a linear programming problem. This makes both classical and restless bandit problems amenable to efficient numerical computations and can also yield some qualitative insight (Nino-Mora [46]).[10] The fourth approach (and the one we emulate) is based on a direct investigation of the sample paths of optimal strategies and an evaluation of the benefits of investing in the risky arm sooner rather than later. While this interchange argument was originally developed by Berry and Fristedt [7] for classical bandit models, it turns out that the monotonicity assumption on the payoffs is what is needed to make the argument work in the more general setting of restless bandits.

### 2.4 Indexability.

Gittins [20] characterized optimal strategies in classical bandit models by an index that is assigned to each arm of the bandit at each instant of time. The optimal strategy is to always choose the arm with the highest index. The indices can be calculated for each arm separately, which reduces the complexity of multi-armed bandits to that of two-armed bandits with one safe and one risky arm.

In general, optimal strategies in restless bandit models do not admit an index representation. Nevertheless, a Lagrangian relaxation of the problem proposed by Whittle [68] yields index strategies that are approximately optimal (Weber and Weiss [65, 64]). The corresponding "Whittle index" (Whittle [68]) is the Lagrange multiplier in a constrained optimization problem and has an economic interpretation as a subsidy for passivity or a fair charge for operating the arm. A major challenge to the deployment of Whittle's index is that it can only be defined when a certain indexability condition is met. In this condition, each arm of the restless bandit is compared to a hypothetical arm with known and constant reward. The indexability condition holds if the set of states where the safe arm is optimal is increasing in the reward from the safe arm.[11]

The question of indexability of restless bandit models is subtle and not yet fully understood. Gittins, Glazebrook, and Weber [21] give an overview of various approaches to establish the indexability of restless bandit models. Partial answers are known for bandits with finite or countable state spaces. Indexability of such models can be tested numerically in a linear programming reformulation of the Markov decision problem (Klimov [32]). In another line of research, Nino-Mora [46] showed that indexability holds for restless bandits satisfying a partial conservation law, which can be verified by running an algorithm. While this can be used to test the indexability of specific restless bandit problems, it does not provide much qualitative insight into which restless bandits are indexable. One would like to have conditions that identify general classes of indexable restless

[10] Another numerical approach is dynamic programming/value function iteration.

[11] This is a monotonicity condition on the optimal strategy, which is not to be confounded with our monotonicity condition on the payoffs and the evolution of human capital.

bandit models – this is the subject of this paper.[12]

# 3 Stochastic control with partial observations.

Section 3.1 provides the general setup. The control problem is formulated in Sections 3.2–3.3. Section 3.4 contains all assumptions and Section 3.5 the main result. Some general notation can be found in Appendix A in the Appendix.

## 3.1 Setup.

$\mathbb{X} = \{0, 1\}$, $\mathbb{Y}$ is a finite dimensional vector space, and $\mathbb{U}$ is a finite set.[13] The hidden state is modeled by an $\mathbb{X}$-valued random variable $X$. The observations are modeled by a càdlàg $\mathbb{Y}$-valued process $Y$. The rewards at time $t$ are given by $b(U_t, X, Y_t)$ for some measurable function $b\colon \mathbb{U} \times \mathbb{X} \times \mathbb{Y} \to \mathbb{R}$. Rewards are discounted exponentially at rate $\rho > 0$ and the aim is to maximize expected discounted rewards.

The evolution of $Y$ depends on a càglàd $\mathbb{U}$-valued process $U$ and on the hidden state $X$. More specifically, the joint distribution of $X$ and $Y$ will be characterized by a controlled martingale problem associated to a linear operator

$$\mathcal{A}\colon \mathcal{D}(\mathcal{A}) \subseteq B(\mathbb{X} \times \mathbb{Y}) \to B(\mathbb{U} \times \mathbb{X} \times \mathbb{Y}).$$

The posterior probability that $X = 1$ given $\{\mathcal{F}_t^Y\}$ is denoted by $P$, i.e., $P$ is a $[0,1]$-valued càdlàg version of the martingale $\mathbb{E}[X \mid \mathcal{F}_t^Y]$. Mathematically speaking, $P$ is called filter of $X$, and economically speaking, the agent's belief in $X = 1$. The joint evolution of $(P, Y)$ will be characterized by a linear operator

$$\mathcal{G}\colon \mathcal{D}(\mathcal{G}) \subseteq B([0,1] \times \mathbb{Y}) \to B(\mathbb{U} \times [0,1] \times \mathbb{Y}).$$

More specific assumptions on $\mathcal{A}$, $\mathcal{G}$, and the payoff function $b$ will be made in Section 3.4.

## 3.2 Control problem with partial observations.

**Definition 1** (Martingale problem for $(\mathcal{A}, F)$)**.** *Let $F$ be a càglàd adapted $\mathbb{U}$-valued process on Skorokhod space $D_{\mathbb{Y}}[0,\infty)$ with its natural filtration. $(X, Y)$ is a solution of the martingale problem for $(\mathcal{A}, F)$ if there exists a filtration $\{\mathcal{F}_t\}$, such that $X$ is an $\mathcal{F}_0$-measurable $\mathbb{X}$-valued random*

[12] Some results in this direction have been obtained for various bandit models related to sensor management, see the survey of Washburn [63]. Other classes of indexable problems are the dual speed problem of Glazebrook, Nino-Mora, and Ansell [23], the maintenance models of Glazebrook, Ruiz-Hernandez, and Kirkbride [24], and the spinning plates and squad models of Glazebrook, Kirkbride, and Ruiz-Hernandez [22]. The spinning plates model is most similar to ours. It satisfies the same monotonicity condition as our model, but has a different reward structure and assumes perfect information.

[13] Our proofs can be generalized to finite state spaces $\mathbb{X}$ at the cost of heavier notation and to compact control spaces $\mathbb{U}$ at the cost of additional criteria ensuring the existence of optimal non-relaxed controls for the discretized separated problem (see e.g. the discussion after Theorem 1.21 in Seierstad [58]).

*variable, $Y$ is an $\{\mathcal{F}_t\}$-adapted càdlàg $\mathbb{Y}$-valued process, and for each $f \in \mathcal{D}(\mathcal{A})$,*

$$f(X, Y_t) - f(X, Y_0) - \int_0^t \mathcal{A}f(F(Y)_s, X, Y_s)\mathrm{d}s$$

*is an $\{\mathcal{F}_t\}$-martingale. The martingale problem is called well-posed if existence and local uniqueness holds under the conditions $X = x$ and $Y_0 = y$, for all $x \in \mathbb{X}$ and $y \in \mathbb{Y}$.*[14]

**Definition 2** (Control with partial observations)**.** *A tuple $(U, X, Y)$ is called a control with partial observations if $U = F(Y)$ holds for some process $F$ as in Definition 1, the martingale problem for $(\mathcal{A}, F)$ is well-posed, and $(X, Y)$ solves the martingale problem for $(\mathcal{A}, F)$.*

**Definition 3** (Value of controls with partial observations)**.** *The value of a control $(U, X, Y)$ with partial observations is defined as*

$$J^{p.o.}(U, X, Y) = \mathbb{E}\left[\int_0^\infty \rho e^{-\rho t} b(U_t, X, Y_t)\mathrm{d}t\right].$$

*The set of controls with partial observations satisfying $\mathbb{E}[X] = p$ and $Y_0 = y$ is denoted by $\mathfrak{C}^{p.o.}_{p,y}$. The value function for the control problem with partial observations is*

$$V^{p.o.}(p, y) = \sup\left\{J^{p.o.}(U, X, Y) : (U, X, Y) \in \mathfrak{C}^{p.o.}_{p,y}\right\}.$$

**Remark 1** (Well-posedness condition)**.** Every càglàd $\{\mathcal{F}^Y_t\}$-adapted process $U$ coincides up to a null set with $F(Y)$ for some process $F$ as in Definition 1 (see Delzeith [13]). Well-posedness of the martingale problem for $(\mathcal{A}, F)$ is, however, a much stronger condition. From the agent's perspective, it requires the control to uniquely determine the outcome. From a mathematical perspective, it excludes pathological cases like the one presented in Example 1 below. It also ensures that controls can be approximated in value by piecewise constant controls, where such degeneracies cannot occur (see Theorem 1).

**Example 1** (Degeneracy in continuous time)**.** Let $\mathbb{X} = \mathbb{U} = \{0, 1\}$, $\mathbb{Y} = \mathbb{R}$, $\mathcal{A}f(u, x, y) = u(2x - 1)f_y(x, y) + \frac{1}{2}uf_{yy}(x, y)$ for each $f \in \mathcal{D}(A) = C^2_b(\mathbb{X} \times \mathbb{Y})$. The aim is to maximize $\mathbb{E}[\int_0^t \rho e^{-\rho t}\mathrm{d}Y_t] = \mathbb{E}[\int_0^\infty \rho e^{-\rho t} b(U_t, X, Y_t)\mathrm{d}t]$ over controls $(U, X, Y)$ of the problem with partial observations, where $b(u, x, y) = u(2x - 1)$. The following tuple $(U, X, Y)$ satisfies all conditions of Definition 2 except for the well-posedness condition: $X$ is a Bernoulli variable, $W$ is Brownian motion independent of $X$, $Y_t = (t + W_t)X$, $U_t = \mathbb{1}_{(0,\infty)}(t)X$, $F(Y)_t = \mathbb{1}_{(0,\infty)}([Y, Y]_t)$. Nevertheless, $U$ depends on the supposedly unobservable state $X$. Actually, $(U, X, Y)$ is optimal for the control problem with observable $X$, and should not be admitted as a control for the problem with unobservable $X$.

**Remark 2** (Topology on the set of controls)**.** So far, there is no topology on the set of controls with partial observations. To get existence of optimal controls, one typically relaxes the control problem

[14] For reference, existence and local uniqueness of the above martingale problem are defined in Appendix B in the Appendix.

by allowing measure-valued controls and shows that the resulting set of admissible controls is compact under some weak topology [36, 16]. In control problems with partial observations involving strong admissibility conditions as in Definition 2, the difficulty is that the set of admissible controls is not weakly closed. This difficulty can be avoided by transforming the problem into a standard problem with full observations, i.e., the separated problem.

### 3.3 Separated control problem.

**Definition 4** (Separated controls)**.** *A tuple $(U, P, Y)$ is called a separated control if there exists a filtration $\{\mathcal{F}_t\}$ such that $U$ is an adapted, càglàd $\mathbb{U}$-valued process, $(P, Y)$ is an adapted, càdlàg $[0,1] \times \mathbb{Y}$-valued process, and for each $f \in \mathcal{D}(\mathcal{G})$, the following process is an $\{\mathcal{F}_t\}$-martingale:*

$$f(P_t, Y_t) - f(P_0, Y_0) - \int_0^t \mathcal{G}f(U_s, P_s, Y_s)\mathrm{d}s.$$

**Definition 5** (Value of separated controls)**.** *The value of a separated control $(U, P, Y)$ is*

$$J^{se.}(U, P, Y) = \mathbb{E}\left[\int_0^\infty \rho e^{-\rho t}\overline{b}(U_t, P_t, Y_t)\mathrm{d}t\right],$$

*where $\overline{b}(u, p, y) = pb(u, 1, y) + (1-p)b(u, 0, y)$. The set of controls $\mathfrak{C}^{se.}_{p,y}$ and the value function $V^{se.}(p, y)$ are defined similarly as in Definition 2.*

**Remark 3** (Filtered martingale problem)**.** Following Stockbridge [60], one could try the alternative approach of defining separated controls as solutions of the filtered martingale problem for $\mathcal{A}$, i.e., the process

$$\Pi_t(\mathrm{d}x) = P_t\delta_1(\mathrm{d}x) + (1 - P_t)\delta_0(\mathrm{d}x)$$

is $\{\mathcal{F}^Y_t\}$-adapted and for each $f \in \mathcal{D}(\mathcal{A})$, the process

$$\int_{\mathbb{X}} f(x, Y_t)\Pi_t(\mathrm{d}x) - \int_{\mathbb{X}} f(x, Y_0)\Pi_0(\mathrm{d}x) - \int_0^t \int_{\mathbb{X}} \mathcal{A}f(U_s, x, Y_s)\Pi_s(\mathrm{d}x)\mathrm{d}s$$

is a martingale with respect to some filtration containing $\{\mathcal{F}^Y_t\}$. Unfortunately, this definition does not rule out the pathological control presented in Example 1, and cannot be used for this reason.

**Remark 4** (Topology on the set of controls)**.** The set of separated controls can be topologized by regarding them as probability measures on the canonical space $L_{\mathbb{U}}[0, \infty) \times D_{[0,1]\times\mathbb{Y}}[0, \infty)$, subject to the condition that the coordinate process solves the martingale problem in Definition 4. Compactness and existence of optimal controls can be obtained by relaxing the control problem. This amounts to replacing $L_{\mathbb{U}}[0, \infty)$ by the space of measures on $\mathbb{U} \times [0, \infty)$ with $[0, \infty)$-marginal equal to the Lebesgue measure and endowing this space with the vague topology [25, 16]. It should be noted, however, that relaxed separated controls are not filters of relaxed controls with partial observations (see Appendix C in the Appendix). In other words, filtering is a non-linear operation on control problems, which does not commute with relaxation.

### 3.4 Specification of the generators and assumptions.

We fix a truncation function $\chi\colon \mathbb{Y} \to \mathbb{Y}$, which is bounded, continuous, and coincides with the identity on a neighborhood of zero.

**Assumption 1** (Operator $\mathcal{A}$)**.** $\mathcal{D}(A) = C_b^2(\mathbb{X} \times \mathbb{Y})$ *and*

$$\begin{aligned}\mathcal{A}f(u,x,y) = \partial_y f(x,y)\beta(u,x,y) + \frac{1}{2}\partial_y^2 f(x,y)\sigma^2(u,y) \\ + \int_{\mathbb{Y}} \Big(f(x,y+z) - f(x,y) - \partial_y f(x,y)\chi(z)\Big)K(u,x,y,\mathrm{d}z),\end{aligned}$$

*where* $\beta\colon \mathbb{U}\times\mathbb{X}\times\mathbb{Y} \to \mathbb{Y}$, $\sigma^2\colon \mathbb{U}\times\mathbb{Y} \to \mathbb{Y}\otimes\mathbb{Y}$, *and* $K$ *is a transition kernel from* $\mathbb{U}\times\mathbb{X}\times\mathbb{Y}$ *to* $\mathbb{Y}\setminus\{0\}$.

**Assumption 2** (Boundedness)**.** *The expressions*

$$b(u,x,y), \qquad \beta(u,x,y), \qquad \sigma^2(u,y), \qquad \int_{\mathbb{Y}} \big(|z|^2 \wedge 1\big)K(u,x,y,\mathrm{d}z)$$

*are bounded over* $(u,x,y) \in \mathbb{U}\times\mathbb{X}\times\mathbb{Y}$.

**Assumption 3** (Girsanov)**.** *There exist functions* $\phi_1\colon \mathbb{U}\times\mathbb{Y} \to \mathbb{Y}$ *and* $\phi_2\colon \mathbb{U}\times\mathbb{Y}\times\mathbb{Y} \to \mathbb{R}$ *satisfying*

$$\begin{aligned}\sigma^2(u,y)\phi_1(u,y) &= \beta(u,1,y) - \beta(u,0,y) - \int_{\mathbb{R}} \big(\phi_2(u,y,z)-1\big)\chi(z)\big(K(u,1,y,\mathrm{d}z) + K(u,0,y,\mathrm{d}z)\big), \\ \phi_2(u,y,z) &= \frac{K(u,1,y,\mathrm{d}z)}{\big(K(u,1,y,\mathrm{d}z) + K(u,0,y,\mathrm{d}z)\big)/2}.\end{aligned}$$

**Assumption 4** (Operator $\mathcal{G}$)**.** $\mathcal{D}(G) = C_b^2([0,1]\times\mathbb{Y})$ *and*

$$\begin{aligned}\mathcal{G}f(u,p,y) = \partial_y f(p,y)\overline{\beta}(u,p,y) + \frac{1}{2}\partial_p^2 f(p,y)p^2(1-p)^2\phi_1(u,y)^\top\sigma^2(u,y)\phi_1(u,y) \\ + \partial_p\partial_y f(p,y)p(1-p)\sigma^2(u,y)\phi_1(u,y) + \frac{1}{2}\partial_y^2 f(p,y)\sigma^2(y,u) \\ + \int_{\mathbb{Y}} \Big(f(p+j(u,p,y,z),y+z) - f(p,y) \\ - \partial_p f(p,y)j(u,p,y,z) - \partial_y f(p,y)\chi(z)\Big)\overline{K}(u,p,y,\mathrm{d}z),\end{aligned}$$

*where*

$$\begin{aligned}\overline{\beta}(u,p,y) &= p\beta(u,1,y) + (1-p)\beta(u,0,y), \\ \overline{K}(u,p,y,\mathrm{d}z) &= pK(u,1,y,\mathrm{d}z) + (1-p)K(u,0,y,\mathrm{d}z), \\ j(u,p,y,z) &= \frac{p\phi_2(u,y,z)}{p\phi_2(u,y,z) + (1-p)\big(2-\phi_2(u,y,z)\big)} - p,\end{aligned}$$

*and where it is understood that* $j(u,p,y,z) = 0$ *if* $p \in \{0,1\}$.

**Assumption 5** (Novikov-style condition)**.** *The following expression is bounded in* $(y,u) \in \mathbb{Y} \times \mathbb{U}$*:*

$$\Phi(u,y) = \frac{1}{8}\phi_1(u,y)^\top \sigma^2(u,y)\phi_1(u,y) + \int_{\mathbb{Y}} \left(1 - \sqrt{\phi_2(u,y,z)\big(2-\phi_2(u,y,z)\big)}\right)\big(K(u,1,y,\mathrm{d}z) + K(u,0,y,\mathrm{d}z)\big).$$

**Assumption 6** (Continuity)**.** *The expressions*

$$\beta(u,x,y), \qquad \sigma^2(u,y), \qquad \phi_1(u,y), \qquad \int_{\mathbb{Y}} g(j(u,p,y,z),z)\overline{K}(u,p,y,\mathrm{d}z)$$

*are continuous in* $(y,u)$ *for all* $x \in \mathbb{X}$*,* $p \in [0,1]$*, and* $g \in C_b([0,1]\times\mathbb{Y})$ *satisfying* $g(x) = O(|x|^2)$ *as* $|x| \to 0$*.*

**Assumption 7** (Condition on big jumps)**.**

$$\lim_{a\to\infty} \sup\Big\{K\big(u,x,y,\{z\in\mathbb{Y}\colon |z|>a\}\big)\colon (u,x,y)\in\mathbb{U}\times\mathbb{X}\times\mathbb{Y}\Big\} = 0.$$

**Assumption 8** (Well-posedness for the problem with partial observations)**.** *The martingale problem for* $(\mathcal{A},F)$ *is well-posed for all deterministic functions* $F\colon [0,\infty)\to\mathbb{U}$*.*

**Assumption 9** (Well-posedness for the separated problem)**.** *The martingale problem for* $(\mathcal{G},u)$ *is well-posed*[15] *for all* $u\in\mathbb{U}$*.*

**Remark 5.**

- The structure of the operator $\mathcal{A}$ in Assumption 1 allows $Y$ to be a general Markovian semimartingale, whereas $X$ is constant. The formula of $\mathcal{G}$ in Assumption 4 comes from Lemma 1, where it is shown that the filter satisfies a martingale problem associated to $\mathcal{G}$.
- The bounds in Assumption 2 guarantee in a simple way that the value functions are finite and reduce technicalities in the proofs.
- The existence of functions $\phi_1,\phi_2$ in Assumption 3 is related to Girsanov's theorem applied to the conditional law of $Y$ given $X$. Assumption 5 is based on a condition in Lépingle and Mémin [40, Théorème IV.3], which establishes uniform integrability of a stochastic exponential. The condition has also an information-theoretic interpretation, see Remark 8.
- Assumptions 6–9 guarantee existence of solutions of martingale problems related to $\mathcal{A}$ and $\mathcal{G}$, and continuous dependence of these solutions on parameters (see e.g. the proof of Lemma 2).

### 3.5 Separation and approximation result

**Theorem 1** (Separation and approximation)**.** *The following statements hold under Assumptions 1–9:*

[15] The martingale problem for $(\mathcal{G},F)$ is defined in analogy to the one for $(\mathcal{A},F)$, see Appendix B in the Appendix.

*(a) The value functions of the control problems agree:*

$$V(p,y) := V^{p.o.}(p,y) = V^{se.}(p,y) < \infty.$$

*(b) Controls can be approximated arbitrarily well in value by piecewise constant controls:*

$$V(p,y) = \sup_{\delta>0} V^{\delta}(p,y),$$

*where $V^{\delta}(p,y) = V^{p.o.,\delta}(p,y) = V^{se.,\delta}(p,y)$ is the value function obtained by restricting to control process $U$ which are piecewise constant on a uniform time grid of step size $\delta > 0$.*

**Remark 6.** The importance of Theorem 1 lies in its capacity to transform the control problem with partial observations into a problem which can be analyzed and solved by standard methods like dynamic programming or linear programming. The approximation result guarantees that the class of admissible strategies is small enough to exclude degeneracies like Example 1. It is also large enough to guarantee the existence of optimal strategies in the restless bandit problem presented in Section 4. In the general case, existence of optimal strategies can be guaranteed by the standard technique of allowing relaxed (measure-valued) controls, as described in Remark 4.

Theorem 1 follows from a sequence of lemmas, which can be found in Appendix D in the Appendix. We now give a verbal proof of the theorem, highlighting the role that each individual lemma plays.

*Proof of Theorem 1.* By Assumption 2 the reward function $b$ is bounded, which implies that all value functions are finite. If $(U,X,Y)$ is a control with partial observations and $P$ is a càdlàg version of the martingale $\mathbb{E}[X \mid \mathcal{F}_t^Y]$, then $(U,P,Y)$ is a separated control with the same value by Lemma 1. Taking the supremum over all controls or step controls, one obtains that

$$V^{\mathrm{p.o.}}(p,y) \le V^{\mathrm{se.}}(p,y), \qquad\qquad V^{\mathrm{p.o.},\delta}(p,y) \le V^{\mathrm{se.},\delta}(p,y).$$

By Lemma 2, separated controls can be approximated arbitrarily well in value by separated step controls. Formally, this is expressed by the equation

$$\sup_{\delta>0} V^{\mathrm{se.},\delta}(p,y) = V^{\mathrm{se.}}(p,y).$$

In Lemma 3, it is shown that Markovian step controls of the separated problem can be transformed into controls of the problem with partial observations of the same value. This is done by a recursive construction, stitching together solutions $(X,Y)$ of the martingale problem associated to $\mathcal{A}$ under constant controls corresponding to each step of the control process. As optimal Markovian controls exist for the discretized separated problem,

$$V^{\mathrm{se.},\delta}(p,y) \le V^{\mathrm{p.o.},\delta}(p,y).$$

Taken together, this implies that

$$V^{\text{se.},\delta}(p,y) = V^{\text{p.o.},\delta}(p,y)$$

and

$$V^{\text{p.o.}}(p,y) \le V^{\text{se.}}(p,y) = \sup_\delta V^{\text{se.},\delta}(p,y) = \sup_\delta V^{\text{p.o.},\delta}(p,y) \le V^{\text{p.o.}}(p,y). \qquad \square$$

# 4 A restless bandit model.

The restless bandit model is formulated in Section 4.1 and solved in Section 4.2. Implications about asymptotics of the filter and strategy are presented in Sections 4.3–4.6.

## 4.1 Setup and assumptions.

The bandit model has a "safe" arm with constant payoffs and a "risky" arm with stochastic payoffs depending on the unobserved "type" $X$ of the agent and her level of "human capital" $H$, which is determined by the amount of past investment in the risky arm. The only departure from Lévy bandits (see Cohen and Solan [10]) is that the Lévy increments of the risky payoff process depend on the investment history.

The general framework of Section 3, including Assumptions 1–9, remains in place. The following structural assumption encodes that (a) the observation process $Y = (H, R)$ takes values in $\mathbb{H}\times\mathbb{R} = \mathbb{R}^2$, (b) the process $H$ has deterministic increments depending only on $U$ and $H$, (c) under a choice $U = 0$ of the safe arm, the reward process $R$ has constant increments, and (d) under a choice $U = 1$ of the risky arm, the reward process $R$ has stochastic increments depending on $X$ and $H$.

**Assumption 10** (Structural assumption)**.** $\mathbb{U} = \{0,1\}$, $\mathbb{Y} = \mathbb{H}\times\mathbb{R} = \mathbb{R}^2$, $Y = (H,R)$. *The coefficients $(\beta,\sigma,K)$ of the generator $\mathcal{A}$ in Assumption 1 are of the form*

$$\beta(1,x,h,r) = \begin{pmatrix}\beta_H(1,h)\\ \beta_R(x,h)\end{pmatrix}, \qquad \beta(0,x,h,r) = \begin{pmatrix}\beta_H(0,h)\\ k\end{pmatrix},$$

$$\sigma^2(1,h,r) = \begin{pmatrix}0 & 0\\ 0 & \sigma_R^2(h)\end{pmatrix}, \qquad \sigma^2(0,h,r) = \begin{pmatrix}0 & 0\\ 0 & 0\end{pmatrix},$$

$$K(1,x,h,r,\mathrm{d}h,\mathrm{d}r) = \delta_0(\mathrm{d}h)K_R(x,h,\mathrm{d}r), \quad K(0,x,h,r,\mathrm{d}h,\mathrm{d}r) = 0,$$

*where*

$$k\in\mathbb{R}, \qquad \beta_H\colon \mathbb{U}\times\mathbb{H}\to\mathbb{H}, \qquad \beta_R\colon \mathbb{X}\times\mathbb{H}\to\mathbb{R}, \qquad \sigma_R\colon \mathbb{H}\to\mathbb{R},$$

*and $K_R$ is a transition kernel from $\mathbb{X}\times\mathbb{H}$ to $\mathbb{R}\setminus\{0\}$ satisfying $\sup_{x,h}\int_{\mathbb{R}}|r|^2\wedge|r|K_R(x,h,\mathrm{d}r)<\infty$.*

In line with the literature on Lévy bandits, the rewards received at time $t$ are given by the

infinitesimal increment $\mathrm{d}R_t$, i.e., the aim is to maximize

$$\mathbb{E}\left[\int_0^\infty \rho e^{-\rho t}\mathrm{d}R_t\right] = \mathbb{E}\left[\int_0^\infty \rho e^{-\rho t} b(U_t, X, H_t)\mathrm{d}t\right],$$

where the reward function $b$ is given as in Assumption 11 below (see Lemma 4).

**Assumption 11** (Reward function)**.** *The reward function $b\colon \mathbb{U}\times\mathbb{X}\times\mathbb{H}$ is given by*

$$b(u,x,h) = \begin{cases} \beta_R(x,h) + \int_{\mathbb{R}} \big(r-\chi(r)\big) K_R(x,h,\mathrm{d}r), & \text{if } u=1,\\ k, & \text{if } u=0. \end{cases}$$

By the following assumption, investment in the risky arm makes future investments in the risky arm more profitable. This dependence is mediated by the process $H$, which increases with investment in the risky arm and decreases otherwise.

**Assumption 12** (Monotonicity condition)**.** *The condition $\beta_H(0,h) \le 0 \le \beta_H(1,h)$ holds for all $h\in\mathbb{H}$. Moreover, the reward $b(1,x,h)$ of the risky arm is non-decreasing in $x\in\mathbb{X}$ and $h\in\mathbb{H}$.*

## 4.2 Reduction to optimal stopping.

**Definition 6** (Gittins' index)**.** *Gittins' index $G$ is defined by*[16]

$$G(p,h) = \inf\left\{ s : \sup_T \mathbb{E}\left(\int_0^T \rho e^{-\rho t}(\mathrm{d}R_t - s\mathrm{d}t)\right) \le 0\right\} = \sup_T \frac{\mathbb{E}\left(\int_0^T \rho e^{-\rho t}\mathrm{d}R_t\right)}{\mathbb{E}\left(\int_0^T \rho e^{-\rho t}\mathrm{d}t\right)},$$

*where $(1,P,H,R)$ is a separated control with constant control process $U\equiv 1$ and initial condition $(P_0,H_0)=(p,h)$, and where the suprema are taken over all $\{\mathcal{F}_t^{P,H}\}$-stopping times $T$.*

**Theorem 2** (Optimal stopping)**.** *The following statements hold under Assumptions 1–12.*

*(a) The value function $V$ (see Theorem 1) does not depend on the initial value of the process $R$ and can be written as $V=V(p,h)$.*

*(b) The strategy $U_t^* = \mathbb{1}_{[\![0,T^*]\!]}(t)$ is optimal, where*

$$T^* = \inf\{t\ge 0 : V(P_t,H_t)\le k\} = \inf\{t\ge 0 : G(P_t,H_t)\le k\}.$$

**Remark 7.**

- The intuition behind Theorem 2 is that choosing the risky arm early rather than late has two advantages: first, it reveals useful information about the hidden state $X$ early on, and second, it makes future rewards from the risky arm more profitable without depreciating rewards from the safe arm.

[16] The index does not depend on the initial value of $R$ (see Lemma 5). The two expressions for $G$ in Definition 6 are shown to be equivalent in El Karoui and Karatzas [15].

- The elimination of the state variable $r$ is possible because of Assumption 10, which asserts that the drift, volatility, and jump measure of the reward process only depend on $P$ and $H$.
- At the heart of Theorem 2 lies the assertion that any optimal control of the discretized problem can be transformed into a stopping rule of at least the same value (Lemma 8). The argument is based on Berry and Fristedt [7, Theorem 5.2.2], but in our setting rewards may depend on the history of experimentation with the risky arm. This dependence is subject to the monotonicity properties in Assumption 12. Our proof shows that these properties are exactly what is needed to adapt the argument of Berry and Fristedt [7] to a restless bandit setting.
- The strategy $U^*$ is well-defined and optimal for the separated problem as well as the problem with partial observations.

Theorem 2 follows from a sequence of lemmas, which can be found in Appendix E in the Appendix. The following proof explains the rôle that each individual lemma plays.

*Proof of Theorem 2.* The value function does not depend on the initial value of $R$ by Lemma 5. Therefore, it can be written as $V(p,h)$. The discrete-time value function $V^\delta(p,h)$ is non-decreasing in $(p,h)$ and convex in $p$. This is established in Lemma 6 using the monotonicity properties in Assumption 12. The result is used in Lemma 7 to prove a sufficient condition for the optimality of the risky arm in the discretized problem: if the myopic payoff is higher for the risky than for the safe arm, then choosing the risky arm is uniquely optimal. This sufficient condition is used in Lemma 8 to prove that $V^\delta(p,h)$ is a supremum of values of stopping rules. The approximation result of Theorem 1 implies that $V(p,y)$ is also a supremum of values of stopping rules. The stopping time $T^* = \inf\{t \geq 0 \colon V(P_t, H_t) \leq k\}$ is optimal by Lemma 9. The alternative characterization of $T^*$ in terms of Gittins' index is well-known, see e.g. Morimoto [45, Theorem 2.1] or El Karoui and Karatzas [15, Proposition 3.4]. □

An immediate consequence of Theorem 2 is a characterization of optimal strategies by a curve which is typically called the *decision frontier*.

**Proposition 1** (Decision frontier)**.** *There is a curve in the $(p,h)$-domain such that it is optimal to invest in the risky arm if $(P_t, H_t)$ lies to the right and above of the curve. Otherwise, it is optimal to invest in the safe arm.*

*Proof.* The value function $V(p,h)$ is non-decreasing in its arguments by Lemma 6 and bounded from below by the constant $k$. The desired curve is the boundary of the domain $\{(p,h) : V(p,h) > k\}$. The characterization of optimal strategies via the position of $(P_t, H_t)$ relative to the curve follows from Theorem 2. □

### 4.3 Indexability.

Another consequence of Theorem 2 is the indexability of our restless bandit model in the sense of Whittle [68].

**Definition 7** (Indexability). *Consider a two-armed bandit problem with a safe and a risky arm. The bandit problem is called indexable if the set of states where the safe arm is optimal is increasing in the payoff $k$ of the safe arm.*

**Proposition 2** (Indexability). *The restless bandit model of Section 4.1 is indexable.*

*Proof.* Gittins' index $G(p,h)$ depends only on the payoff of the risky arm. Therefore, the set $\{(p,h)\colon G(p,h) \leq k\}$ where the safe arm is optimal has the required properties. □

### 4.4 Asymptotic learning.

**Definition 8** (Asymptotic learning and infinite investment). *For any $\omega \in \Omega$, we say that asymptotic learning holds if $\lim_{t\to\infty} P_t(\omega) = X(\omega)$. We say that the agent invests an infinite amount of time in the risky arm if $\int_0^\infty U_t(\omega)\mathrm{d}t = \infty$.*

**Assumption 13** (Bounds on the flow of information). *The initial belief is non-doctrinaire, i.e., $P_0 \in (0,1)$. The measures $K_R(1,h,\cdot)$ and $K_R(0,h,\cdot)$ are equivalent, for all $h \in \mathbb{H}$. The function $\Phi(1,\cdot)$ defined in Assumption 2 is bounded from below by a positive constant.*

**Proposition 3** (Asymptotic learning). *Under Assumptions 1–13, the following statements hold:*

*(a) Under any control, asymptotic learning occurs if and only if the agent invests an infinite amount in the risky arm.*

*(b) Under the optimal control of Theorem 2, asymptotic learning takes place if and only if $(P,H)$ remains above the decision frontier for all time.*

*Proof.* (a) follows from Lemma 10. (b) follows from (a) and the characterization of optimal controls in Proposition 1. □

**Remark 8.**

- The limit $\lim_{t\to\infty} P_t$ exists almost surely because $P$ is a bounded martingale. If the belief $P_0 \in \{0,1\}$ is doctrinaire, then the belief process $P$ is constant and equal to the hidden state $X$.
- Agents can learn their true type $X$ in two ways: either through a jump of the belief process to $X$, or through convergence to $X$ without a jump to the limit. The first kind of learning is excluded by the equivalence of $K_R(1,h,\cdot)$ and $K_R(0,h,\cdot)$. The second kind of learning is characterized by divergence of the Hellinger process of the measures $\mathbb{P}_1$ and $\mathbb{P}_0$. The Hellinger process is closely related to the function $\Phi(u,y)$, which can be interpreted as the informativeness of the arm $u$ about the state $X$. The upper and lower bounds on $\Phi$ in Assumptions 5 and 13 establish an equivalence between divergence of the Hellinger process and divergence of the accumulated amount of investment in the risky arm (see Lemma 10).

- If the measures $K_R(1, H, \cdot)$ and $K_R(0, H, \cdot)$ are not equivalent, the belief process $P$ jumps to the true state $X$ with positive probability on any finite interval of time where the risky arm is chosen. For example, this is the case in the exponential bandits model of Keller, Rady, and Cripps [31].

- Proposition 3 can be contrasted with the strategic experimentation model of Bolton and Harris [9] and the social learning model of Acemoglu et al. [1, Example 1.1]. In these models, asymptotic learning always takes place because agents continuously receive information about the hidden state, regardless of whether they choose to invest or not.

### 4.5 Comparison to the full-information case.

By the full-information case, we mean the bandit model where the otherwise hidden state variable $X$ is fully observable. This model is equivalent to the model with partial observations and $P_0 \in \{0, 1\}$. It follows from Theorem 2 and the monotonicity condition in Assumption 12 that the optimal strategy in the full-information case is constant in time and given by $\mathbb{1}_{V(X,H_0)>k}$.

**Definition 9** (Asymptotic efficiency)**.** *For any $t \geq 0$ and $\omega \in \Omega$, $U_t(\omega)$ is called efficient if it coincides with $\mathbb{1}_{V(X(\omega),H_0)>k}$. Moreover, $U(\omega)$ is called asymptotically efficient if $U_t(\omega)$ is efficient for all sufficiently large times $t$.*

**Assumption 14** (Decision frontier stays away from $p = 0$ and $p = 1$)**.** *There is $\epsilon > 0$ such that for all $h \in \mathbb{H}$, $V(\epsilon, h) = k$ and $V(1, h) > k$.*

**Proposition 4** (Asymptotic efficiency)**.** *Let Assumptions 1–14 hold, let $U$ be the optimal strategy provided by Theorem 2, and assume that $(P_0, H_0)$ lies above the decision frontier. Conditional on $X = 0$, asymptotic efficiency holds almost surely. Conditional on $X = 1$, however, asymptotic efficiency may hold and fail with positive probability.*

*Proof.* If $X = 0$, investment in the risky arm can't continue forever. Otherwise, $P_t$ would converge to zero by Proposition 3. As the decision frontier is strictly bounded away from the set $p = 0$, $(P_t, H_t)$ would eventually drop below the decision frontier, a contradiction. Thus, investment stops at some finite point in time. This is efficient given $X = 0$ because $V(0, H_0) = k$.

If $X = 1$, then $(P, H)$ may or may not drop below the frontier at some point in time. Both cases may happen with positive probability. In the former case, the agent stops investing, which is inefficient because $V(1, H_0) > 0$. In the latter case, the agent never stops investing, which is efficient. □

**Remark 9.**

- Efficiency holds if there is some time $t$ where the agent's plan for future investments is the same as if she had known $X$ from the beginning. Of course, this still leaves open the possibility that some early investment decisions were inefficient.

- The intuition behind Proposition 4 is that a sequence of bad payoffs can lead agents to refrain from experimentation with the risky arm. For agents of the type $X = 0$, this is efficient, but

for agents with $X = 1$, it is not. In this regard, the restless bandit model behaves as a standard bandit model.

- It follows that in the long run, compared to a setting with full information, agents invest too little in the risky arm. This points to the importance of policies designed to increase investment in the risky arm.

- Assumption 14 limits the influence of $H$ on the rewards from the risky arm: the safe arm is optimal if $X = 0$ is known for sure, regardless of how high $H$ is, and similarly the risky arm is optimal if $X = 1$ is known for sure, regardless of how low $H$ is.

- Without Assumption 14, it is still possible to characterize asymptotic efficiency using the necessary and sufficient conditions of Lemma 10, but there are more cases to distinguish. Some of them have no counterpart in classical bandit models. For example, there can be low-type agents who invest in the risky arm at all times. This can be either efficient or inefficient, depending on whether $V(0, H_0)$ exceeds $k$. Similarly, it can be efficient or inefficient for high-type agents to stop investing, depending on $V(1, H_0)$.

### 4.6 Evolution of a population of agents.

Assume that there is a population of agents with idiosyncratic $(P_0, H_0)$. Moreover, assume that agents have independent types, such that learning from others is impossible. Alternatively, learning could be precluded by making actions and rewards private information. Then all agents behave as in the single player case. The distribution of agents in the $(p, h)$-domain evolves over time and converges to the distribution of $(P_\infty, H_\infty)$.

**Proposition 5.** *Let Assumptions 1–14 hold, let $p^*(h)$ denote the decision frontier, and consider a population of agents with $(P_0, H_0)$ above the decision frontier.*

*(a) In a restless bandit model with $\beta(0, h) < 0 < \beta(1, h)$, $(P_\infty, H_\infty)$ satisfies*

$$P_\infty \in [0, p^*(-\infty)] \text{ and } H_\infty = -\infty \qquad \text{or} \qquad P_\infty = 1 \text{ and } H_\infty = \infty.$$

*(b) In a classical bandit model where $\beta(0, h) = \beta(1, h) = 0$ and $\Delta P \geq -\epsilon$ for some $\epsilon \geq 0$,*

$$P_\infty \in [p^*(H_0) - \epsilon, p^*(H_0)] \text{ and } H_\infty = H_0.$$

*Proof.* (a) If $(P, H)$ drops below the decision frontier, $P$ is frozen and $H$ decreases to $-\infty$. Otherwise, $P$ increases to 1 and $H$ to $\infty$. (b) $H$ is constant and $P$ either converges to 1 or drops below the decision frontier and remains there forever. □

**Remark 10.**

- Proposition 5 shows that agents in classical bandit models accumulate at or near the decision frontier, whereas they drift away from the frontier in restless bandit models.

- It follows that in restless bandit models, even large shifts of the decision frontier have negligible effects on average investment if they are carried out late in time. In contrast, even small shifts of the decision frontier have large effects on average investment in classical bandit models, where agents accumulate near the frontier.

- Our model provides an explanation for the ineffectiveness of subsidies designed to boost investment in projects with uncertain payoffs. Our explanation does not rely on switching costs.

# 5 Conclusions.

We presented an extension of classical bandit models of investment under uncertainty motivated by dynamic aspects of resource development. The extension is new and has economic significance in a wide range of real world settings.

We dealt with the delicate issue of setting up the control problem with partial observations in continuous time. As explained in Section 2.2, recent standard formulations of optimal control under partial observation do not apply in our general setting. In addition to its importance to the theory of optimal control, our solution is also a contribution to the bandit literature.

Our framework encompasses both the exponential bandit model of Keller, Rady, and Cripps [31], where jumps can occur only for high type agents, and the Poisson and Levy bandit models of Keller and Rady [30, 29] and Cohen and Solan [10], where it is assumed that one jump measure is absolutely continuous with respect to the other.

We solved the restless bandit model by an unconventional approach. Instead of using the HJB equation or a setup using time changes, we discretized the problem in time and showed that any optimal strategy can be modified such that the agent never invests after a period of not investing and such that the modified strategy is still optimal.

Our models constitute a new class of indexable restless bandit models. While other classes of indexable bandits are known, they either involve no learning about one's type (Glazebrook, Kirkbride, and Ruiz-Hernandez [22]), do not allow history-dependent payoffs (Washburn [63]), or are restricted to very specific reward processes (e.g. finite-state Markov chains as in Nino-Mora [46]).

# A Notation.

For any Polish space $\mathbb{S}$, $B(\mathbb{S})$ will denote the space of $\mathbb{R}$-valued Borel-measurable functions on $\mathbb{S}$, $C(\mathbb{S})$ the continuous functions, $C_b(\mathbb{S})$ the bounded continuous functions, and $\mathcal{P}(\mathbb{S})$ the space of probability measures on $\mathbb{S}$. $D_{\mathbb{S}}[0,\infty)$ denotes the space of $\mathbb{S}$-valued càdlàg functions on $[0,\infty)$ with the Skorokhod topology, $L_{\mathbb{S}}[0,\infty)$ the càglàd functions, and $C_{\mathbb{S}}[0,\infty)$ the subspace of continuous functions. If $\mathbb{S}$ is endowed with a differentiable structure, then $C_b^k(\mathbb{S})$ denotes the functions with $k$ bounded continuous derivatives.

Throughout the paper, all filtrations are assumed to be complete, and all processes are assumed to be progressively measurable. The law of a random variable $X$ is denoted by $\mathcal{L}(X)$. The completion of the filtration generated by a process $Y$ is denoted by $\{\mathcal{F}_t^Y\}$. If $Y$ has left-limits, they are denoted by $Y_-$, i.e., $Y_{t-} = \lim_{s \nearrow t} Y_t$. If $Y$ is of finite variation, $\mathrm{Var}(Y)$ denotes its variation process. $H \bullet Y$ denotes stochastic integration of a predictable process $H$ with respect to a semimartingale $Y$ and $H * \mu$ with respect to a random measure $\mu$. $I$ denotes the identity process $I_t = t$. When $T$ is a stopping time, we write $Y^T$ and $\mu^T$ for the stopped versions of $Y$ and $\mu$. Stochastic intervals are denoted by double brackets, e.g., $[\![0, T]\!] \subset [0, \infty] \times \Omega$. $Y^c$ denotes the continuous local martingale part of $Y$. A superscript $\top$ denotes the transpose of a matrix or vector.

# B Controlled martingale problems.

**Definition 10** (Martingale problem for $(\mathcal{A}, F)$)**.** *Let $F$ be a càglàd adapted $\mathbb{U}$-valued process on the space $D_{\mathbb{Y}}[0, \infty)$ with its canonical filtration.*

*(i) $(X, Y, T)$ is a solution of the stopped martingale problem for $(\mathcal{A}, F)$ if there exists a filtration $\{\mathcal{F}_t\}$, such that $X$ is an $\mathcal{F}_0$-measurable $\mathbb{X}$-valued random variable, $Y$ is an $\{\mathcal{F}_t\}$-adapted càdlàg $\mathbb{Y}$-valued process, $T$ is an $\{\mathcal{F}_t\}$-stopping time, and for each $f \in \mathcal{D}(\mathcal{A})$,*

$$f(X, Y_{t\wedge T}) - f(X, Y_0) - \int_0^{t\wedge T} \mathcal{A}f(F(Y)_s, X, Y_s)\mathrm{d}s \tag{1}$$

*is an $\{\mathcal{F}_t\}$-martingale.*

*(ii) If $T = \infty$ almost surely, then $(X, Y)$ is a solution of the martingale problem for $(\mathcal{A}, F)$.*

*(iii) $(X, Y)$ is a solution of the local martingale problem for $(\mathcal{A}, F)$ if there exists a filtration $\{\mathcal{F}_t\}$ and a sequence of $\{\mathcal{F}_t\}$-stopping times $\{T_n\}$ such that $T_n \to \infty$ almost surely and for each $n$, $(X, Y, T_n)$ is a solution of the stopped martingale problem for $(\mathcal{A}, F)$.*

*(iv) Local uniqueness holds for the martingale problem for $(\mathcal{A}, F)$ if for any solutions $(X', Y', T')$, $(X'', Y'', T'')$ of the stopped martingale problem for $(\mathcal{A}, F)$, equality of the law of $(X', Y_0')$ and $(X'', Y_0'')$ implies the existence of a solution $(X, Y, S' \vee S'')$ of the stopped martingale problem for $(\mathcal{A}, F)$ such that $(X_{\cdot\wedge S'}, S')$ has the same distribution as $(X'_{\cdot\wedge T'}, T')$, and $(X_{\cdot\wedge S''}, S'')$ has the same distribution as $(X''_{\cdot\wedge T''}, T'')$.*

*(v) The martingale problem for $(\mathcal{A}, F)$ is well-posed if local uniqueness holds for the martingale problem for $(\mathcal{A}, F)$ and for each $\nu \in \mathcal{P}(\mathbb{X} \times \mathbb{Y})$, there exists a solution $(X, Y)$ of the local martingale problem for $(\mathcal{A}, F)$ such that the law of $(X, Y_0)$ is $\nu$.*

**Definition 11** (Martingale problem for $(\mathcal{G}, F)$)**.** *Let $F$ be a càglàd adapted $\mathbb{U}$-valued process on $D_{[0,1]\times\mathbb{Y}}[0, \infty)$ with its canonical filtration.*

*(i)* $(P, Y, T)$ *is a solution of the stopped martingale problem for* $(\mathcal{G}, F)$ *if there exists a filtration* $\{\mathcal{F}_t\}$*, such that* $(P, Y)$ *is an* $\{\mathcal{F}_t\}$*-adapted càdlàg* $[0,1] \times \mathbb{Y}$*-valued process,* $T$ *is an* $\{\mathcal{F}_t\}$*-stopping time, and for each* $f \in \mathcal{D}(\mathcal{G})$*,*

$$f(P_{t\wedge T}, Y_{t\wedge T}) - f(P_0, Y_0) - \int_0^{t\wedge T} \mathcal{A}f(F(P,Y)_s, P_s, Y_s)\mathrm{d}s \tag{2}$$

*is an* $\{\mathcal{F}_t\}$*-martingale.*

*(ii) Solutions of the (local) martingale problem, local uniqueness, and well-posedness are defined in analogy to Definition 10.*

# C Noncommutativity of filtering and relaxation.

To see the non-commutativity between filtering and relaxation, let us tentatively define relaxed controls with partial observations as tuples $(\Lambda, X, Y)$ such that for each $f \in \mathcal{D}(\mathcal{A})$,

$$f(X, Y_t) - f(X_0, Y_0) - \int_0^t \int_{\mathbb{U}} \mathcal{A}f(u, X, Y_{s-})\Lambda_s(\mathrm{d}u)\mathrm{d}s \tag{3}$$

is a martingale, where $\Lambda$ is a $\{\mathcal{F}_t^Y\}$-predictable $\mathcal{P}(\mathbb{U})$-valued process. If a well-posedness condition similar to the one in Definition 2 holds and $P_t = \mathbb{E}[X \mid \mathcal{F}_t^Y]$ is the filter, then it can be shown[17] that a jump $\Delta Y_t$ of the observable process leads to a jump $\Delta P_t = \bar{j}(\Lambda_t, P_{t-}, Y_{t-}, \Delta Y_t)$ of the filter, where

$$\bar{j}(\lambda, p, y, z) = \frac{\int_{\mathbb{U}} p\phi_2(u, y, z)\lambda(\mathrm{d}u)}{\int_{\mathbb{U}} \Big(p\phi_2(u, y, z) + (1-p)\big(2 - \phi_2(u, y, z)\big)\Big)\lambda(\mathrm{d}u)} - p. \tag{4}$$

Thus, $\Delta P_t$ is uniquely determined by $\Delta Y_t$ and the information before $t$. In contrast, this is not the case in the relaxation of the separated control problem, where a jump $\Delta Y_t$ can lead to different values of $\Delta P_t$. Indeed, the jump measure of $(P, Y)$ is compensated by the predictable random measure

$$\nu(\mathrm{d}p, \mathrm{d}y) = \int_{\mathbb{Y}} \delta_{j(u, P_{t-}, Y_{t-}, y)}(\mathrm{d}p)\overline{K}(u, P_{t-}, Y_{t-}, \mathrm{d}y)\Lambda_t(\mathrm{d}u). \tag{5}$$

An interpretation is that the two cases differ in how uncertainty regarding $u$ is handled. In the former case, the control $u$ in the support of $\Lambda_t$ is treated as unknown in the process of updating the filter. Therefore, the jump height of the filter depends on $\Lambda_t$, but not on a random choice of $u$ in the support of $\Lambda_t$. In the latter case, however, $u$ is treated as known but random. Different choices of $u$ in the support of $\Lambda_t$ might lead to different probabilities for a jump $\Delta Y_t$, and consequently to different jumps of the filter.

[17] This follows by adapting the proof of Lemma 1 to relaxed control processes.

# D Proofs of Section 3.

Lemmas 1–3 below are used to establish Theorem 1. Assumptions 1–9 are in place.

**Lemma 1** (Filtering)**.** *If $(U, X, Y)$ is a control with partial observations and $P$ is a càdlàg version of the martingale $\mathbb{E}[X \mid \mathcal{F}_t^Y]$, then $(U, P, Y)$ is a separated control of the same value as $(U, X, Y)$.*

*Proof. Step 1 (Filter as change of measure from $\mathbb{P}$ to $\mathbb{P}_1$).* If $P_0 \in \{0, 1\}$, then $P_t \equiv P_0$ is constant and equal to $X$. In this case it is trivial to check that $(U, P, Y)$ is a separated control of the same value as $(U, X, Y)$. In the sequel, we assume that $0 < P_0 < 1$. Then the measure $\mathbb{P}$ can be conditioned on the event $X = x$, for all $x \in \mathbb{X}$. This yields measures $\mathbb{P}_x$ such that

$$\mathbb{P}_1(X = 1) = 1, \qquad \mathbb{P}_0(X = 0) = 1, \qquad \mathbb{P} = P_0\mathbb{P}_1 + (1 - P_0)\mathbb{P}_0. \tag{6}$$

The process $P/P_0$ is the $\{\mathcal{F}_t^Y\}$-density process of $\mathbb{P}_1$ relative to $\mathbb{P}$ because for all $A \in \mathcal{F}_t^Y$,

$$\int_A P_t \mathrm{d}\mathbb{P} = \int_A \mathbb{E}[X|\mathcal{F}_t^Y]\mathrm{d}\mathbb{P} = \int_A X \mathrm{d}\mathbb{P} = P_0\mathbb{P}_1(A). \tag{7}$$

*Step 2 (Stochastic exponential relating the martingale problems under $\mathbb{P}$ and $\mathbb{P}_1$).* For each $f \in \mathcal{D}(\mathcal{A})$, let $\overline{\mathcal{A}}f$ be the average of $\mathcal{A}f$ over $x \in \mathbb{X}$ with weights $p$ and $(1 - p)$,

$$\overline{\mathcal{A}}f(u, p, y) = p\mathcal{A}f(u, 1, y) + (1 - p)\mathcal{A}f(u, 0, y). \tag{8}$$

Let $f \in \mathcal{D}(\mathcal{A})$ and set $g(x, y) = f(1, y)$. Then $g \in \mathcal{D}(A)$ and $g$ is constant in $x \in \mathbb{X}$. By Definition 2, the process

$$g(1, Y) - g(1, Y_0) - \mathcal{A}g(U, X, Y) \bullet I \tag{9}$$

is a martingale under $\mathbb{P}$. Taking $\{\mathcal{F}_t^Y\}$-optional projections, one obtains that the process

$$M = g(1, Y) - g(1, Y_0) - \overline{\mathcal{A}}g(U, P, Y) \bullet I \tag{10}$$

is an $\{\mathcal{F}_t^Y\}$-martingale under $\mathbb{P}$. Moreover, as $X = 1$ holds $\mathbb{P}_1$-a.s., the process

$$\widetilde{M} = g(1, Y) - g(1, Y_0) - \mathcal{A}g(U, 1, Y) \bullet I \tag{11}$$

is an $\{\mathcal{F}_t^Y\}$-martingale under $\mathbb{P}_1$. The difference between these two processes is given by

$$\begin{aligned} M - \widetilde{M} = {} & \partial_y g(1, Y)\big(\beta(U, 1, Y) - \overline{\beta}(U, P, Y)\big) \bullet I \\ & + \int_{\mathbb{Y}} \big(g(1, Y + z) - g(1, Y) - \partial_y g(1, Y)\chi(z)\big)\big(K(U, 1, Y, \mathrm{d}z) - \overline{K}(U, P, Y, \mathrm{d}z)\big) \bullet I. \end{aligned} \tag{12}$$

For any $p > 0$, let $\psi_1$ and $\psi_2$ be defined by

$$\psi_1(u, p, y) = (1 - p)\phi_1(u, y), \quad \psi_2(u, p, y, z) = \frac{\phi_2(u, y, z)}{p\phi_2(u, y, z) + (1 - p)\big(2 - \phi_2(u, y, z)\big)}, \tag{13}$$

where $\phi_1, \phi_2$ stem from Assumption 5. Then the following relations hold for any $p > 0$:

$$\begin{aligned} \beta(u,1,y) - \overline{\beta}(u,p,y) &= \sigma^2(u,y)\psi_1(u,p,y) + \int_{\mathbb{R}^n} \big(\psi_2(u,p,y,z) - 1\big)\chi(z)\overline{K}(u,p,y,\mathrm{d}z), \\ K(u,1,y,\mathrm{d}z) &= \psi_2(u,p,y,z)\overline{K}(u,p,y,\mathrm{d}z) \end{aligned} \tag{14}$$

For any $n \in \mathbb{N}$, let $T_n$ be the stopping time

$$T_n = \inf\{t \geq 0 \colon P_t < 1/n \text{ or } P_{t-} < 1/n \text{ or } |Y_t| > n\} \wedge n. \tag{15}$$

Since $P > 0$ holds on any interval $[\![0, T_n[\![$, Equation (14) can be used to rewrite Equation (12) as

$$\begin{aligned} M^{T_n} - \widetilde{M}^{T_n} = \partial_y g(1,Y)\big(\sigma^2(U,Y)\psi_1(U,P,Y)\big)\mathbb{1}_{[\![0,T_n]\!]} \bullet I \\ + \int_{\mathbb{Y}} \big(g(1,Y+z) - g(1,Y)\big)\big(\psi_2(U,P,Y,z) - 1\big)\overline{K}(U,P,Y,\mathrm{d}z)\mathbb{1}_{[\![0,T_n]\!]} \bullet I. \end{aligned} \tag{16}$$

Let $\mu$ be the jump measure of $Y$, $\nu$ its $\{\mathcal{F}_t^Y\}$-compensator under $\mathbb{P}$, and

$$L^n = \psi_1(U,P_-,Y_-)\mathbb{1}_{[\![0,T_n]\!]} \bullet Y^c + \big(\psi_2(U,P_-,Y_-,z) - 1\big)\mathbb{1}_{[\![0,T_n]\!]} * \big(\mu - \nu\big)(\mathrm{d}z,\mathrm{d}t). \tag{17}$$

Then $L^n$ is a local $\{\mathcal{F}_t^Y\}$-martingale under $\mathbb{P}$. Keeping track of the terms in Itō's formula the same way as in the proof of Jacod and Shiryaev [26, Theorem II.2.42] shows that

$$M = \partial_y g(1,Y_-) \bullet Y^c + \big(g(1,Y_- + z) - g(1,Y_-)\big) * \big(\mu - \nu\big)(\mathrm{d}z,\mathrm{d}t) \tag{18}$$

is the decomposition of $M$ into its continuous and purely discontinuous local martingale parts. It is now easy to calculate the predictable quadratic covariation of $M^{T_n}$ and $L^n$. Indeed, a comparison with Equation (16) shows that

$$M^{T_n} - \widetilde{M}^{T_n} = \langle M^{T_n}, L^n \rangle. \tag{19}$$

Equivalently, letting $D^n = \mathcal{E}(L^n)$ denote the stochastic exponential of $L^n$,

$$\widetilde{M}^{T_n} = M^{T_n} - \langle M^{T_n}, L^n \rangle = M^{T_n} - \frac{1}{D^n_-} \bullet \langle M^{T_n}, D^n \rangle. \tag{20}$$

*Step 3 (Martingale property of stochastic exponential).* We will show that the local martingale $D^n$ is a martingale by verifying the conditions of Lépingle and Mémin [40, Théorème IV.3]. For any $w \in [0,1]$,

$$\frac{p(1-p)}{pw + (1-p)(1-w)} \leq \frac{p(1-p)}{p \wedge (1-p)} \leq 1 \tag{21}$$

holds because the nominator on the left-hand side is a convex combination of $p$ and $(1-p)$. Replacing

$w$ by $\phi_2(u,y,z)/2$ in Equation (21) one obtains

$$\big(\psi_2(u,p,y,z)-1\big)^2 = \left(\frac{2(1-p)(\phi_2(u,y,z)-1)}{p\phi_2(u,y,z)+(1-p)\big(2-\phi_2(u,y,z)\big)}\right)^2 \le \frac{1}{p^2}\big(\phi_2(u,y,z)-1\big)^2. \tag{22}$$

This inequality relates the values of $\phi_2, \psi_2$ under the transformation $w \mapsto (w-1)^2$. It can equivalently be expressed in terms of the functions $w \mapsto w\log(w) - w + 1$ or $w \mapsto 1 - \sqrt{w(2-w)}$ because for all $w \in [0,2]$,

$$w\log(w) - w + 1 \le (w-1)^2 \le 4\big(w\log(w) - w + 1\big), \tag{23}$$

$$1 - \sqrt{w(2-w)} \le (w-1)^2 \le 2\left(1 - \sqrt{w(2-w)}\right). \tag{24}$$

Actually, the first inequality in Equation (23) holds for all $w \ge 0$, which implies that

$$\begin{aligned}
&\int \Big(\psi_2(u,p,y,z)\log\big(\psi_2(u,p,y,z)\big) - \psi_2(u,p,y,z) + 1\Big)\overline{K}(u,p,y,\mathrm{d}z) \\
&\quad\le \int \big(\psi_2(u,p,y,z)-1\big)^2\overline{K}(u,p,y,\mathrm{d}z) \le \frac{1}{p^2}\int \big(\phi_2(u,y,z)-1\big)^2\overline{K}(u,p,y,\mathrm{d}z) \\
&\quad\le \frac{2}{p^2}\int \left(1 - \sqrt{\phi_2(u,y,z)\big(2-\phi_2(u,y,z)\big)}\right)\overline{K}(u,p,y,\mathrm{d}z) \\
&\quad\le \frac{4}{p^2}\int \left(1 - \sqrt{\phi_2(u,y,z)\big(2-\phi_2(u,y,z)\big)}\right)\overline{K}(u,\tfrac{1}{2},y,\mathrm{d}z).
\end{aligned} \tag{25}$$

By Assumption 5, this expression is bounded as long as $p$ stays away from zero. Moreover, by the same assumption, the following expression is bounded:

$$\psi_1(u,p,y)^\top \sigma^2(u,y)\psi_1(u,p,y) = (1-p)^2\phi_1(u,y)^\top\sigma^2(u,y)\phi_1(u,y). \tag{26}$$

Therefore,

$$\mathbb{E}\left[\exp\left(\frac{1}{2}\langle L^{n,c}, L^{n,c}\rangle_\infty + \big((1+z)\log(1+z) - z\big) * \nu_\infty^{T_n}\right)\right] < \infty, \tag{27}$$

which is the condition of Lépingle and Mémin [40, Théorème IV.3] implying that $D^n = \mathcal{E}(L^n)$ is a uniformly integrable martingale. Therefore, $D^n_{T_n}\mathbb{P}$ is a probability measure.

*Step 4 (Identification of stochastic exponential and filter).* By Definition 2, $U = F(Y)$ for a process $F$ on $D_{\mathbb{Y}}[0,\infty)$. We will use the well-posedness of the martingale problem for $(\mathcal{A}, F)$ to show that $D^n_{T_n}\mathbb{P}$ agrees with $\mathbb{P}_1$ on $\mathcal{F}^Y_{T_n}$. By Girsanovs' theorem and Equation (20), $\widetilde{M}^{T_n}$ is an $\{\mathcal{F}^Y_t\}$-martingale under $D^n_{T_n}\mathbb{P}$. The process $\widetilde{M}$ can be written as

$$\widetilde{M} = f(1,Y) - f(1,Y_0) - \mathcal{A}f(U,1,Y) \bullet I \tag{28}$$

because $\mathcal{A}$ has no derivatives or non-local terms in the $x$-direction. As $f \in \mathcal{D}(\mathcal{A})$ was chosen arbitrarily, the tuple $(1,Y)$ under the measure $D^n_{T_n}$ solves the martingale problem for $(\mathcal{A}, F)$ stopped at

$T_n$. The same can be said about the tuple $(X, Y)$ under the measure $\mathbb{P}_1$. Moreover, the distribution of $(1, Y_0)$ under $D^n_{T_n}\mathbb{P}$ coincides with the distribution of $(X, Y_0)$ under $\mathbb{P}_1$. According to Definition 2, local uniqueness holds for the martingale problem. It follows that $D^n_{T_n}\mathbb{P}$ coincides with $\mathbb{P}_1$ on $\mathcal{F}^Y_{T_n}$. The characterization of $P/P_0$ as the density process of the measure $\mathbb{P}_1$ relative to $\mathbb{P}$ obtained in Step 1 implies that $P = P_0 D^n_{T_n}$ holds on $[\![0, T_n]\!]$.

*Step 5 (Filter solves the martingale problem with generator $\mathcal{G}$).* To show that $(U, P, Y)$ is a separated control, one has to prove that for any $f \in \mathcal{D}(\mathcal{G})$,

$$N = f(P, Y) - f(P_0, Y_0) - \mathcal{G}f(U, P, Y) \bullet I$$

is an $\{\mathcal{F}^Y_t\}$-martingale under $\mathbb{P}$. On the interval $[\![0, T_n]\!]$, $P$ agrees with $P_0 D^n$ and consequently satisfies $P = P_- \bullet L^n$. Therefore, the jumps of $P$ on this interval are

$$\Delta P = P_-\, \Delta L^n = P_-\big(\psi_2(U, P_-, Y_-, \Delta Y) - 1\big)\mathbb{1}_{\Delta Y \neq 0} = j(U, P_-, Y_-, \Delta Y)\mathbb{1}_{\Delta Y \neq 0}, \tag{29}$$

where the function $j$ is defined in Assumption 4. Moreover, on the same interval $[\![0, T_n]\!]$,

$$\begin{aligned}
\langle P^c, P^c\rangle &= P_-^2 \bullet \langle L^{n,c}, L^{n,c}\rangle = \sum_{i,j} P_-^2 \psi_{1,i}(U, P_-, Y_-)\psi_{1,j}(U, P_-, Y_-) \bullet \langle Y^{i,c}, Y^{j,c}\rangle \\
&= P^2(1-P)^2\phi_1(U, Y)^\top \sigma(U, Y)^2 \phi_1(U, Y) \bullet I \\
\langle P^c, Y^c\rangle &= P_-^2 \bullet \langle L^{n,c}, Y^c\rangle = P_-^2 \psi_1(U, P_-, Y_-)^\top \bullet \langle Y^c, Y^c\rangle = P^2 \psi_1(U, P, Y)^\top \sigma^2(U, Y) \bullet I \\
\langle Y^c, Y^c\rangle &= \sigma^2(U, Y) \bullet I.
\end{aligned} \tag{30}$$

It follows from Itō's formula and the definition of $\mathcal{G}$ in Assumption 4 that the stopped process $N^{T_n}$ is an $\{\mathcal{F}^Y_t\}$-local martingale under $\mathbb{P}$. It is also bounded by Assumption 2, so it is a martingale. Setting $g(x, y) = f(0, y)$, one has $g \in \mathcal{D}(\mathcal{A})$, and the process

$$M = g(0, Y) - g(0, Y_0) - \overline{\mathcal{A}}g(U, P, Y) \bullet I \tag{31}$$

is a martingale. Then it holds for any bounded stopping time $S$ and each $n \in \mathbb{N}$ that

$$\mathbb{E}\big[N_S\big] = \mathbb{E}\big[N_{S\wedge T_n} + N_{S\vee T_n} - N_{T_n}\big] = \mathbb{E}\big[N_{S\wedge T_n} + M_{S\vee T_n} - M_{T_n} + R_n\big] = \mathbb{E}\big[R_n\big], \tag{32}$$

with a remainder $R_n$ given by

$$R_n = \big(N_{S\vee T_n} - N_{T_n}\big) - \big(M_{S\vee T_n} - M_{T_n}\big). \tag{33}$$

Let $\omega \in \Omega$ and

$$T = \lim_{n\to\infty} T_n = \inf\{t \geq 0 \colon P_t = 0 \text{ or } P_{t-} = 0\}. \tag{34}$$

If $P_{T-}(\omega) = 0$, then $T_n(\omega) < T(\omega)$ holds for all $n \in \mathbb{N}$. Otherwise, there is $k \in \mathbb{N}$ such that

$T_n(\omega) = T(\omega)$ holds for all sufficiently large $n$. Therefore,

$$\lim_{n\to\infty} R_n = \begin{cases} (N_{T-} - N_{T-}) - (M_{T-} - M_{T-}), & \text{if } P_{T-} = 0 \text{ and } t < T, \\ (N_S - N_{T-}) - (M_S - M_{T-}), & \text{if } P_{T-} = 0 \text{ and } t \geq T, \\ (N_{S\vee T} - N_T) - (M_{S\vee T} - M_T), & \text{if } P_{T-} \neq 0. \end{cases} \tag{35}$$

It can be seen from the definitions of $\mathcal{A}$ and $\mathcal{G}$ in Assumptions 1 and 4 that $\mathcal{G}f(u,0,y) = \overline{\mathcal{A}}g(u,0,y)$. Therefore, $N = M$ holds on the interval $[\![T,\infty[\![$, where $P = 0$. Moreover, $N_{T-} = M_{T-}$ holds if $P_{T-} = 0$. This implies that $\lim_{n\to\infty} R_n = 0$. The processes $M^S$ and $N^S$ are bounded, which follows from Assumption 2 and the boundedness of $S$. Therefore,

$$R_n = \mathbb{1}_{T_n<S}\big((N_S - N_{S\wedge T_n}) - (M_S - M_{S\wedge T_n})\big) \tag{36}$$

is bounded by a constant not depending on $n$. By the dominated convergence theorem, $\mathbb{E}[N_S] = \lim_{n\to\infty} \mathbb{E}[R_n] = 0$. As this holds for all bounded stopping times $S$, we conclude that $N$ is a martingale. As $f \in \mathcal{D}(\mathcal{G})$ was chosen freely, $(U,P,Y)$ is a separated control.

*Step 6 (Value of separated control).* $(U,P,Y)$ has the same value as $(U,X,Y)$ because $\overline{b}(U,P,Y)$ is the $\{\mathcal{F}_t^Y\}$-optional projection of $b(U,X,Y)$. □

**Lemma 2** (Approximation)**.** *Separated controls can be approximated arbitrarily well in value by separated step controls:*

$$V^{se.}(p,y) = \sup_\delta V^{se.,\delta}(p,y). \tag{37}$$

*Here $V^{se.,\delta}$ denotes the value function obtained by admitting only processes $U$ which are piecewise constant on an equidistant time grid of step size $\delta > 0$ in the separated control problem.*

*Proof. Step 1 (Filtered martingale problem).* Let $U$ be deterministic and let $(P,Y)$ be a càdlàg process with values in $[0,1]\times\mathbb{Y}$. We identify $P$ with the $\mathcal{P}(\mathbb{X})$-valued process $\Pi$ given by

$$\Pi_t(\mathrm{d}x) = P_t\delta_1(\mathrm{d}x) + (1-P_t)\delta_0(\mathrm{d}x), \tag{38}$$

where $\delta_x$ denotes the Dirac measure at $x \in \mathbb{X}$. In line with Kurtz and Ocone [34], we say that $(\Pi,Y)$ is a solution of the filtered martingale problem for $(\mathcal{A},U)$ if

$$\int_{\mathbb{X}} f(x,Y)\Pi(\mathrm{d}x) - \int_{\mathbb{X}} f(x,Y_0)\Pi_0(\mathrm{d}x) - \int_{\mathbb{X}} \mathcal{A}f(U,x,Y)\Pi(\mathrm{d}x) \bullet I \tag{39}$$

is a martingale, for each $f \in \mathcal{D}(\mathcal{A})$, and $\Pi$ is $\{\mathcal{F}_t^Y\}$-adapted.

We will use Kurtz and Nappo [37, Theorem 3.6] to show that uniqueness holds for the filtered martingale problem. Thus, we have to verify points (i)-(vi) of Condition 2.1 in this paper. These are conditions on the operator $\mathcal{A}f(U,x,y)$ in (39), interpreted as a time-dependent generator of $(X,Y)$. To put everything into a time-homogeneous framework, we work with the time-augmented

process $(I, X, Y)$. Its generator $\mathcal{A}^U$ is given by

$$\mathcal{D}(\mathcal{A}^U) = C_b^2(\mathbb{R} \times \mathbb{X} \times \mathbb{Y}), \qquad \mathcal{A}^U g(t,x,y) = \partial_t g(t,x,y) + \mathcal{A} g_t(U_t, x, y), \tag{40}$$

where $g_t(x,y) = g(t,x,y)$. For point (i), there is nothing to prove. For point (ii), one has to show that $\mathcal{A}f(u,x,y)$ is continuous in $(u,x,y)$, for each $f \in \mathcal{D}(\mathcal{A})$. To see this, let

$$m(u,x,y) = 1 + \int_{\mathbb{R}^d} \big(|z|^2 \wedge 1\big) K(u,x,y,\mathrm{d}z), \tag{41}$$

$$\hat{f}(u,x,y,z) = m(u,x,y) \frac{f(x,y+z) - f(x,y) - \partial_y f(x,y)\chi(z)}{|z|^2 \wedge 1}, \tag{42}$$

$$\hat{K}(u,x,y,\mathrm{d}z) = \frac{\big(|z|^2 \wedge 1\big) K(u,x,y,\mathrm{d}z)}{m(u,x,y)}. \tag{43}$$

Then everything is set up such that

$$\int_{\mathbb{Y}} \big(f(x,y+z) - f(x,y) - \partial_y f(x,y)\chi(z)\big) K(u,x,y,\mathrm{d}z) = \int_{\mathbb{Y}} \hat{f}(u,x,y,z) \hat{K}(u,x,y,\mathrm{d}z). \tag{44}$$

Now let $(u_n, x_n, y_n)_{n\in\mathbb{N}}$ be a sequence in $\mathbb{U} \times \mathbb{X} \times \mathbb{Y}$ converging to $(u,x,y)$. By Assumption 6, $m(u,x,y)$ is continuous and the measures $\hat{K}(u_n, x_n, y_n, \mathrm{d}z)$ are weakly convergent. A version of Skorokhod's representation theorem for measures instead of probability measures (for example Startek [59]) implies that there are mappings $(Z_n)_{n\in\mathbb{N}}$ and $Z$ with values in $\mathbb{Y}$, all defined on the same measure space with finite measure, such that for each $n \in \mathbb{N}$, $Z_n$ has distribution $\hat{K}(u_n, x_n, y_n, \mathrm{d}z)$, $Z$ has distribution $\hat{K}(u,x,y,\mathrm{d}z)$, and $Z_n \to Z$ almost surely. By the dominated convergence theorem,

$$\mathbb{E}\big[\hat{f}(u_n, x_n, y_n, Z_n)\big] \to \mathbb{E}\big[\hat{f}(u,x,y,Z)\big], \tag{45}$$

which shows that the expression in (44) is continuous in $(u,x,y)$. This settles point (ii). Point (iii) is satisfied with $\psi = 1$ by Assumption 2. Points (iv) and (vi) are satisfied for $\mathcal{D}(\mathcal{A}^U) = C_b^2(\mathbb{R}\times\mathbb{X}\times\mathbb{Y})$. Finally, point (v) is satisfied because of Assumption 8, which guarantees that for each constant, deterministic control $U$ and all initial conditions, there exists a càdlàg solution of the martingale problem for $\mathcal{A}^U$ (cf. the discussion before Theorem 2.1 in Kurtz [35]). Moreover, by Assumption 8, uniqueness holds for the martingale problem for $\mathcal{A}^U$, which coincides with the martingale problem for $(\mathcal{A}, U)$ from Definition 1. Thus, all conditions of Kurtz and Nappo [37, Theorem 3.6] are fulfilled and uniqueness holds for the filtered martingale problem.

*Step 2 (Projecting separated controls to solutions of the filtered martingale problem).* Let $(U, P, X)$ be a separated control with deterministic control process $U$. Let $f \in \mathcal{D}(\mathcal{G})$ be affine in the first variable $p$, i.e.,

$$f(p,x) = pf(1,x) + (1-p)f(0,x). \tag{46}$$

Then $\mathcal{G}f(u,p,x)$ is also affine in $p$, i.e.,

$$\mathcal{G}f(u,p,x) = p\mathcal{G}f(u,0,x) + (1-p)\mathcal{G}f(u,0,x) = p\mathcal{A}f(u,0,x) + (1-p)\mathcal{A}f(u,0,x). \tag{47}$$

This can be verified using the definition of $\phi_1$ and $\phi_2$, noting that all quadratic terms in $p$ cancel out in the expression of $\mathcal{G}f(u,p,x)$. Identifying $P$ with $\Pi$ as in Step 1, one obtains that the process

$$\begin{aligned} &f(P,Y) - f(P_0,Y_0) - \mathcal{G}f(U,P,Y) \bullet I \\ &\qquad = \int_{\mathbb{X}} f(x,Y)\Pi(\mathrm{d}x) - \int_{\mathbb{X}} f(x,Y_0)\Pi_0(\mathrm{d}x) - \int_{\mathbb{X}} \mathcal{A}f(U,x,Y)\Pi(\mathrm{d}x) \bullet I \end{aligned} \tag{48}$$

is a martingale. If $\widetilde{\Pi}$ denotes the $\{\mathcal{F}_t^Y\}$-optional projection of $\Pi$, then

$$\int_{\mathbb{X}} f(x,Y)\widetilde{\Pi}(\mathrm{d}x) - \int_{\mathbb{X}} f(x,Y_0)\widetilde{\Pi}_0(\mathrm{d}x) - \int_{\mathbb{X}} \mathcal{A}f(U,x,Y)\widetilde{\Pi}(\mathrm{d}x) \bullet I \tag{49}$$

is also a martingale. Thus, $(\tilde{\Pi}, Y)$ is a solution of the filtered martingale problem. By the previous step, the law of $(\tilde{\Pi}, Y)$ is uniquely determined. An important consequence is that all separated controls sharing the same deterministic control process $U$ have the same value:

$$\mathbb{E}\left[\int_0^\infty \rho e^{-\rho t}\bar{b}(U_t,P_t,Y_t)\mathrm{d}t\right] = \mathbb{E}\left[\int_0^\infty \rho e^{-\rho t}\bar{b}(U_t,\widetilde{P}_t,Y_t)\mathrm{d}t\right] =: J(U), \tag{50}$$

where $\widetilde{P}$ is the $\{\mathcal{F}_t^Y\}$-optional projection of $P$.

*Step 3 (Tightness of separated controls).* Let $(U^n,P^n,Y^n)$ be separated controls with deterministic control processes $U^n$ and let $U^n \to U$ in the stable topology, i.e.,

$$\int_0^\infty g(U_t^n,t)\mathrm{d}t \to \int_0^\infty g(U_t,t)\mathrm{d}t \tag{51}$$

for all bounded measurable functions $g\colon \mathbb{U}\times[0,\infty) \to \mathbb{R}$ with compact support which are continuous in $u$. The stable topology coincides with the vague topology, checked on continuous functions with compact support. For more details on the vague and stable topology we refer to El Karoui, Nguyen, and Jeanblanc-Picqué [16] and Jacod and Mémin [25]. We will use Jacod and Shiryaev [26, Theorem IX.3.9] to show that the laws of $(P^n,Y^n)$ are tight. Thus, we have to verify the conditions of this theorem. By the same estimates as in Step 3 of the proof of Lemma 1, one obtains that

$$\begin{aligned} &\int_{\mathbb{Y}} \big(|j(u,p,y,z)|^2 + |z|^2\big) \wedge 1 \ \overline{K}(u,p,y,\mathrm{d}z) \\ &\quad \le 2\int_{\mathbb{Y}} j(u,p,y,z)^2 \ \overline{K}(u,p,y,\mathrm{d}z) + 2\int_{\mathbb{Y}} |z|^2 \wedge 1 \ \overline{K}(u,p,y,\mathrm{d}z) \\ &\quad \le 2\int_{\mathbb{Y}} \big(\phi_2(u,y,z)-1\big)^2 \ \overline{K}(u,p,y,\mathrm{d}z) + 2\int_{\mathbb{Y}} |z|^2 \wedge 1 \ \overline{K}(u,p,y,\mathrm{d}z) \\ &\quad \le 4\int_{\mathbb{Y}} \left(1 - \sqrt{\phi_2(u,y,z)\big(2-\phi_2(u,y,z)\big)}\right) \overline{K}(u,p,y,\mathrm{d}z) + 2\int_{\mathbb{Y}} |z|^2 \wedge 1 \ \overline{K}(u,p,y,\mathrm{d}z). \end{aligned} \tag{52}$$

By Assumptions 2 and 5, the integrals on the last line above are bounded by a constant which does not depend on $(u,p,y)$. By Assumptions 2 and 5, also the drift and the diagonal entries of the volatility matrix

$$\overline{\beta}(u,p,y), \qquad p^2(1-p)^2\phi_1(u,y)^\top\sigma^2(u,y)\phi_1(u,y), \qquad \sigma^2(u,y) \tag{53}$$

are bounded by a constant not depending on $(u,p,y)$. It follows that Condition IX.3.6 (the strong majoration hypothesis) is satisfied. Condition IX.3.7 (the condition on the big jumps) follows from Assumption 7. By the stable convergence of $U^n$ to $U$, using Assumption 6 and the bounds which were just shown, the following convergence holds for all $t \geq 0$, $(P,Y) \in D_{[0,1]\times\mathbb{Y}}[0,\infty)$, and functions $g \in C_b([0,1]\times\mathbb{Y})$ vanishing near the origin:

$$\begin{aligned}
\overline{\beta}(U^n,P,Y)\bullet I_t &\to \overline{\beta}(U,P,Y)\bullet I_t,\\
P^2(1-P)^2\phi_1(U^n,Y)^\top\sigma^2(U^n,Y)\phi_1(U^n,Y)\bullet I_t &\to P^2(1-P)^2\phi_1(U,Y)^\top\sigma^2(U,Y)\phi_1(U,Y)\bullet I_t\\
P(1-P)\sigma^2(U^n,Y)\phi_1(U^n,Y)\bullet I_t &\to P(1-P)\sigma^2(U,Y)\phi_1(U,Y)\bullet I_t\\
\sigma^2(U^n,Y)\bullet I_t &\to \sigma^2(U,Y)\bullet I_t,\\
\int_{\mathbb{Y}} g\big(j(U^n,P,Y,z),z\big)\,\overline{K}(U^n,P,Y,\mathrm{d}z)\bullet I_t &\to \int_{\mathbb{Y}} g\big(j(U,P,Y,z),z\big)\,\overline{K}(U,P,Y,\mathrm{d}z)\bullet I_t.
\end{aligned} \tag{54}$$

It follows from Lemma IX.3.4 that the conditions of Theorem IX.3.9 are satisfied. Thus, the laws of $(P^n,Y^n)$ are tight. Moreover, any limit $(P,Y)$ of a weakly converging subsequence of $(P^n,Y^n)$ solves the martingale problem for $(\mathcal{G},U)$ and defines a separated control $(U,P,Y)$. This follows from Jacod and Shiryaev [26, Theorem IX.2.11] by the same assumptions.

*Step 4 (Step controls).* For any $\delta > 0$, the mapping

$$\Psi^\delta\colon L_{\mathbb{U}}[0,\infty) \to L_{\mathbb{U}}[0,\infty), \qquad (\Psi^\delta U)_t = \sum_{i=0}^{\infty} U_{i\delta}\,\mathbb{1}_{(i\delta,(i+1)\delta]}(t) \tag{55}$$

approximates deterministic control processes by step control processes of step size $\delta$. Indeed, $\lim_{\delta\to 0}(\Psi^\delta U)_t = U_t$ holds for each $t \geq 0$. Moreover, by dominated convergence, $\Psi^\delta U$ converges stably to $U$. Let

$$L^0_{\mathbb{U}}[0,\infty) = \bigcup_{\delta>0}\Big\{\Psi^\delta U\colon U \in L_{\mathbb{U}}[0,\infty)\Big\} \subset L_{\mathbb{U}}[0,\infty) \tag{56}$$

denote the set of all step control processes. For any step control process $U \in L^0_{\mathbb{U}}[0,\infty)$, there is a control with partial observations $(U,X,Y)$ by Assumption 8 and a corresponding separated control $(U,P,Y)$ by Lemma 1. Let $\mathbb{Q}_U$ denote the law of $(P,Y)$ under $U$. If $U^n \in L^0_{\mathbb{U}}[0,\infty)$ converges stably to a step control $U \in L^0_{\mathbb{U}}[0,\infty)$, then $\mathbb{Q}_{U^n}$ converges weakly to $\mathbb{Q}_U$ by the arguments in Step 2 and by Assumption 9 ensuring uniqueness of the martingale problem for $(\mathcal{G},U)$. As continuity implies measurability, $\mathbb{Q}$ is a transition kernel from $L^0_{\mathbb{U}}[0,\infty)$ with the Borel sigma algebra of stable convergence to Skorokhod space $D_{[0,1]\times\mathbb{Y}}[0,\infty)$.

*Step 5 (Approximation of deterministic controls).* Let $U \in L_{\mathbb{U}}[0,\infty)$ and define $U^n = \Psi^{1/n}U$, for each $n \in \mathbb{N}$. By Assumption 8 there are controls with partial observations $(U^n, X^n, Y^n)$ and by Lemma 1 corresponding separated controls $(U^n, P^n, Y^n)$. By the tightness result of Step 3, any subsequence along which $J(U^n)$ converges contains another subsequence, still denoted by $n$, such that $(P^n, Y^n)$ converge weakly to some solution $(P, Y)$ of the martingale problem for $(\mathcal{G}, U)$. By Skorokhod's representation theorem we may assume after passing to yet another subsequence that $(P^n, Y^n)$ and $(P, Y)$ are defined on the same probability space and that $(P^n, Y^n)$ converge to $(P, Y)$ almost surely. As $\Delta P_t = 0$ holds almost surely for each fixed $t \geq 0$, it follows from the dominated convergence theorem and the pointwise convergence of $U_t^n$ to $U_t$ that

$$\lim_{n\to\infty} J(U^n) = \int_0^\infty \rho e^{-\rho t} \lim_{n\to\infty} \mathbb{E}\left[\overline{b}(U_t^n, P_t^n, Y_t^n)\right] \mathrm{d}t = \int_0^\infty \rho e^{-\rho t}\, \mathbb{E}\left[\overline{b}(U, P_t, Y_t)\right] \mathrm{d}t = J(U). \quad (57)$$

*Step 5 (Approximation of arbitrary controls).* The law of any separated control $(U, P, Y)$ is a probability measure $\mathbb{P}$ on the space $L_{\mathbb{U}}[0,\infty) \times D_{[0,1]\times\mathbb{Y}}[0,\infty)$. We will work on this canonical probability space in the sequel. Using disintegration, $\mathbb{P}$ can be written in the form

$$\mathbb{P}(\mathrm{d}U, \mathrm{d}P, \mathrm{d}Y) = \mathbb{P}(\mathrm{d}U)\mathbb{P}_U(\mathrm{d}P, \mathrm{d}Y). \quad (58)$$

Accordingly, the value of the control can be expressed as

$$J^{\text{se.}}(U, P, Y) = \int_{L_{\mathbb{U}}[0,\infty)} \mathbb{E}_{\mathbb{P}_U}\left[\int_0^\infty \rho e^{-\rho t}\overline{b}(U_t, P_t, Y_t)\mathrm{d}t\right] \mathbb{P}(\mathrm{d}U) \quad (59)$$

For $\mathbb{P}$-a.e. $U$, the process $(P, Y)$ under the measure $\mathbb{P}_U$ solves the martingale problem for $(\mathcal{G}, U)$. Moreover, the process $U$ is deterministic under the measure $\mathbb{P}_U$. By Step 2, all solutions of the martingale problem $(\mathcal{G}, U)$ with deterministic control process $U$ have the same value $J(U)$. This allows one to express the value of the control as

$$J^{\text{se.}}(U, P, Y) = \int_{L_{\mathbb{U}}[0,\infty)} J(U)\mathbb{P}(\mathrm{d}U). \quad (60)$$

By Step 4 and dominated convergence,

$$J^{\text{se.}}(U, P, Y) = \lim_{n\to\infty} \int_{L_{\mathbb{U}}[0,\infty)} J(\Psi^{1/n}U)\mathbb{P}(\mathrm{d}U) = \lim_{n\to\infty} J^{\text{se.}}(U^n, P^n, Y^n), \quad (61)$$

where $(U^n, P^n, Y^n)$ is the coordinate process on $L_{\mathbb{U}}[0,\infty) \times D_{[0,1]\times\mathbb{Y}}[0,\infty)$ under the measure $\mathbb{Q}_{\Psi^{1/n}U}(\mathrm{d}P, \mathrm{d}Y)\mathbb{P}(\mathrm{d}U)$. Thus, $(U^n, P^n, Y^n)$ is a sequence of separated step controls approximating $(U, P, Y)$ in value. □

**Lemma 3** (From separated to partially observed controls)**.** *For every separated step control, there exists a step control with partial observations of at least the same value, implying* $V^{p.o.,\delta}(p, y) \geq V^{se.,\delta}(p, y)$.

*Proof. Step 1 (Reduction to Markovian step controls).* To distinguish the separated and the partially observed versions of the problem, we will mark objects of the separated problem with a tilde. By Assumption 9, the discretized separated problem is that of controlling the Markov chain $(\widetilde{P}_{t_i}, \widetilde{Y}_{t_i})$, where $(t_i)_{i\in\mathbb{N}}$ is a uniform time grid of step size $\delta > 0$. It is well-known that optimal Markov controls exist for such problems (see e.g. Berry and Fristedt [7] and Seierstad [58]). We will prove the lemma by showing that every Markov control for the discretized, separated problem corresponds to a step control for the problem with partial observations which has the same value. So we start with a Markovian step control $(\widetilde{U}, \widetilde{P}, \widetilde{Y})$ with control process $\widetilde{U}$ given by

$$\widetilde{U}_t = F_i(\widetilde{P}_{t_i}, \widetilde{Y}_{t_i}), \qquad \text{if } t \in (t_i, t_{i+1}], \tag{62}$$

for some functions $F_i \colon [0,1] \times \mathbb{Y} \to \mathbb{U}$, $i \in \mathbb{N}$.

*Step 2 (Construction of a candidate control with partial observations).* To construct the control for the problem with partial observations, we work on the canonical space $\Omega = \mathbb{X} \times D_{\mathbb{Y}}[0,\infty)$ with its natural sigma algebra and filtration. The coordinates on this space are denoted by $(X, Y)$. When $T$ is a (strict) stopping time, $\mathbb{P}$ is a probability measure on $\Omega$, and $\mathbb{Q}$ is an $\mathcal{F}_T$-measurable random variable with values in the space of probability measures on $\Omega$, then we let $\mathbb{P} \otimes_T \mathbb{Q}$ denote the unique probability measure on $\Omega$ such that (i) the law of the stopped process $(X, Y^T)$ is equal to $\mathbb{P}$ on the sigma algebra $\mathcal{F}_T$ and (ii) the $\mathcal{F}_T$-conditional law of the time-shifted process $(X, Y_{T+t})_{t\geq 0}$ is $\mathbb{Q}$. This notation is explained and relevant results are proven in Stroock and Varadhan [61, 6.1.2, 6.1.3 and 1.2.10] for continuous processes. For processes with jumps, the relevant results are Jacod and Shiryaev [26, Lemmas III.2.43-48], but the notation $\mathbb{P} \otimes_T \mathbb{Q}$ is not used there.

By Assumption 8, we get for each $(u, x, y) \in \mathbb{U} \times \mathbb{X} \times \mathbb{Y}$ a unique probability measure $\mathbb{Q}^u(x, y)$ on $\Omega$ such that $X = x$ and $Y_0 = y$ holds almost surely and such that $(X, Y)$ solve the martingale problem for $(\mathcal{A}, u)$ under $\mathbb{Q}^u(x, y)$. By Jacod and Shiryaev [26, Theorem IX.3.39], $\mathbb{Q}^u(x, y)$ is weakly continuous, thus measurable, in $(u, x, y)$. Verifying the conditions of the theorem can be done as in the proof of Lemma 2, but it is easier in the present situation. We now define inductively for each $n \in \mathbb{N}$ a probability measure $\mathbb{P}^n$ and a càdlàg process $P^n$ on $\Omega$ as follows.

$$\begin{aligned} \mathbb{P}^0 &= P_0 \mathbb{Q}^{F_0(P_0,Y_0)}(1, Y_0) + (1 - P_0)\mathbb{Q}^{F_0(P_0,Y_0)}(0, Y_0), & P^0_t &= \mathbb{E}_{\mathbb{P}^0}[X \mid \mathcal{F}^Y_t], \\ \mathbb{P}^n &= \mathbb{P}^{n-1} \otimes_{t_n} \mathbb{Q}^{F_n(P^{n-1}_{t_n}, Y_{t_n})}(X, Y_{t_n}), & P^n_t &= \mathbb{E}_{\mathbb{P}^n}[X \mid \mathcal{F}^Y_t]. \end{aligned} \tag{63}$$

It follows that the measures $\mathbb{P}^n$ and $\mathbb{P}^m$ agree on $\mathcal{F}_{t_n \wedge t_m}$ and that the processes $P^n$ and $P^m$ agree almost surely on $[0, t_n \wedge t_m]$. Therefore, there is a unique measure $\mathbb{P}$ which coincides with $\mathbb{P}^n$ on $\mathcal{F}_{t_n}$, for all $n$. Furthermore, there is a unique càdlàg process $P$ that is almost surely equal to $P^n$ on $[0, t_n]$, for all $n$. If $U$ is defined as

$$U_t = \sum_{i=0}^{\infty} F_i(P_{t_i}, Y_{t_i}) \mathbb{1}_{(t_i, t_{i+1}]}(t), \tag{64}$$

then by construction, the process

$$f(X,Y) - f(X,Y_0) - \mathcal{A}f(U,X,Y) \bullet I \tag{65}$$

is a martingale, for each $f \in \mathcal{D}(\mathcal{A})$ (see also Jacod and Shiryaev [26, Lemma III.2.48]).

*Step 3 (Verification of the well-posedness condition).* As $P$ is the $\{\mathcal{F}_t^Y\}$-optional projection of $X$, it is indistinguishable from $G(Y)$ for some càdlàg process $G$ on $D_{\mathbb{Y}}[0,\infty)$ by Delzeith [13]. It follows from Equation (64) that $U$ is indistinguishable from $F(Y)$ for some càglàd process $F$ on $D_{\mathbb{Y}}[0,\infty)$. The martingale problem $(\mathcal{A}, F)$ is well-posed by Assumption 8 because $F$ is a step process. Thus, the well-posedness condition of Definition 2 is satisfied.

*Step 4 (Value of the control with partial observations).* The process $P$ defined in Step 2 is the $\{\mathcal{F}_t^Y\}$-optional projection of $X$. By Lemma 1, $(U,P,Y)$ defines a separated control of the same value as $(U,X,Y)$. Assumption 9 implies that $(U,P,Y)$ is equal in law to $(\widetilde{U},\widetilde{P},\widetilde{Y})$. Therefore, $(U,X,Y)$ has the same value as $(\widetilde{U},\widetilde{P},\widetilde{Y})$. □

# E Proofs of Section 4.

The setup of Section 4.1, including Assumptions 1–12, holds.

**Lemma 4** (Payoff function)**.** *For any control $(U,X,H,R)$ of the problem with partial observations,*

$$\mathbb{E}\left[\int_0^\infty \rho e^{-\rho t} \mathrm{d}R_t\right] = \mathbb{E}\left[\int_0^\infty \rho e^{-\rho t} b(U_t, X, H_t)\mathrm{d}t\right], \tag{66}$$

*where $b$ is given by Assumption 11.*

*Proof.* By the integrability condition on $K_R$ in Assumption 10, the process $R$ is a special semi-martingale. Its canonical decomposition is

$$R = R_0 + b(U,X,H) \bullet I + R^c + r * (\mu - \nu), \tag{67}$$

where $\mu$ is the integer-valued random measure associated to the jumps of $R$ and $\nu = \mathbb{1}_{U=1} K_R(X,H,\cdot)$ is the compensator of $\mu$. For $\zeta_t = \rho e^{-\rho t}$ one obtains that

$$\zeta \bullet R - \zeta b(U,X,H) \bullet I = \zeta \bullet R^c + \zeta r * (\mu - \nu) \tag{68}$$

is a local martingale. Equation (66) holds if it is a true martingale.

Let $\chi_R(r) = \chi(0,r)$. The processes $\zeta \bullet R^c$ and $\zeta\chi_R(r) * (\mu - \nu)$ are square integrable martingales by the Burkholder-Davis-Gundy inequality because their quadratic variations are integrable:

$$\begin{aligned} \mathbb{E}\big[[\zeta \bullet R^c]_\infty\big] &= \mathbb{E}\left[\zeta^2 \mathbb{1}_{U=1} \sigma_R(H)^2 \bullet I_\infty\right] < \infty, \\ \mathbb{E}\big[[\zeta\chi_R(r) * (\mu-\nu)]_\infty\big] &= \mathbb{E}\big[\zeta^2 \chi_R(r)^2 * \mu_\infty\big] = \mathbb{E}\big[\zeta^2 \chi_R(r)^2 * \nu_\infty\big] < \infty. \end{aligned} \tag{69}$$

This follows from the bounds on $\sigma_R$ and $K_R$ in Assumptions 2 and 10. Furthermore, the process $(r-\chi_R(r))*(\mu-\nu)$ is a uniformly integrable martingale on $[0,t]$ because it is of integrable variation:

$$\begin{aligned}\mathbb{E}\big[\operatorname{Var}(\zeta(r-\chi_R(r))*(\mu-\nu))_\infty\big] &\leq \mathbb{E}\left[\zeta|r-\chi_R(r)|*\mu_\infty\right]+\mathbb{E}\left[\zeta|r-\chi_R(r)|*\nu_\infty\right]\\ &= 2*\mathbb{E}\left[\zeta|r-\chi_R(r)|*\nu_\infty\right]<\infty.\end{aligned} \tag{70}$$

This follows from the bound on $K_R$ in Assumption 10. Therefore, the process in Equation (68) is a martingale, and Equation (66) holds. □

**Lemma 5** (Elimination of the state variable $r$)**.** *The value functions $V(p,h,r)$, $V^\delta(p,h,r)$ do not depend on $r$ and can be written as $V(p,h)$, $V^\delta(p,h)$.*

*Proof.* For any $s\in\mathbb{R}$ and $f\in\mathcal{D}(\mathcal{A})$, let $f_s(x,h,r)=f(x,h,r+s)$. Then $f_s\in\mathcal{D}(\mathcal{A})$, and by Assumption 10, $\mathcal{A}f(u,x,h,r+s)=\mathcal{A}f_s(u,x,h,r)$. If $(U,X,H,R)$ is a control with partial observations, then the equation

$$\begin{aligned}f(X,H,R+s)-f(X,H_0,R_0+s)-\mathcal{A}f(U,X,H,R+s)\bullet I&\\ = f_s(X,H,R)-f_s(X,H_0,R_0)-\mathcal{A}f_s(U,X,H,R)\bullet I&\end{aligned} \tag{71}$$

shows that $(U,X,H,R+s)$ is also a control with partial observations. Moreover, the two controls have the same value. The same argumentation applies to separated controls. □

**Lemma 6.** *The value functions $V(p,h)$ and $V^\delta(p,h)$ are convex, non-decreasing in $p$, and non-decreasing in $h$.*

*Proof.* Recall from Step 5 in the proof of Lemma 2 that the value of any separated control can be written as

$$J^{\text{se.}}(U,P,H,R)=\int_{L_{\mathbb{U}}[0,\infty)}J(U)\mathbb{P}(\mathrm{d}U), \tag{72}$$

where $\mathbb{P}(\mathrm{d}U)$ is the marginal distribution of $U\in L_{\mathbb{U}}[0,\infty)$ and $J(U)$ is the value of a deterministic control process $U$. In the definition of $J(U)$ in Equation (50), $P$ is a martingale and $(U,H)$ are deterministic. Therefore,

$$\begin{aligned}J(U)=\ &\mathbb{E}\left[\int_0^\infty \rho e^{-\rho t}\overline{b}(U_t,P_t,H_t)\mathrm{d}t\right]\\ &= P_0\int_0^\infty \rho e^{-\rho t}b(U_t,1,H_t)\mathrm{d}t+(1-P_0)\int_0^\infty \rho e^{-\rho t}b(U_t,0,H_t)\mathrm{d}t.\end{aligned} \tag{73}$$

This expression is linear in $P_0$ and non-decreasing in $(P_0,H_0)$ by Assumption 12. Taking the supremum over all controls or step controls with fixed initial condition $(P_0,H_0)$, one obtains convexity in $P_0$ and monotonicity in $(P_0,H_0)$. □

**Lemma 7** (Sufficient condition for optimality of the risky arm)**.** *In the discretized separated problem, the risky arm is uniquely optimal as an initial choice if its expected first-stage payoff exceeds the first-stage payoff of the safe arm.*

*Proof.* We fix $\delta > 0$ and only allow control processes which are piecewise constant on the uniform time grid of step size $\delta$. The expected first-stage payoff is denoted by

$$\overline{b}^{\delta}(u,p,h) = \mathbb{E}\left[\int_0^{\delta} \rho e^{-\rho t}\overline{b}(u,P_t,H_t)\mathrm{d}t\right], \tag{74}$$

where $(P,H)$ stems from a separated control with initial condition $(P_0,H_0)=(p,h)$ and constant control process $U_t \equiv u$. By Bellman's principle, optimal initial choices $U_0$ for the discretized separated problem are maximizers of

$$\max_{u\in\mathbb{U}} \ \overline{b}^{\delta}(u,p,h) + e^{-\rho\delta}\,\mathbb{E}\left[V^{\delta}(P_{\delta},H_{\delta}) \;\middle|\; (U_0,P_0,H_0)=(u,p,h)\right]. \tag{75}$$

Thus, the optimal initial choice depends on the sign of the quantity

$$\begin{aligned}\overline{b}^{\delta}(1,p,h) - \overline{b}^{\delta}(0,p,h) + e^{-\rho\delta}\,\mathbb{E}\left[V^{\delta}(P_{\delta},H_{\delta}) \;\middle|\; (U_0,P_0,H_0)=(1,p,h)\right]\\ - e^{-\rho\delta}\,\mathbb{E}\left[V^{\delta}(P_{\delta},H_{\delta}) \;\middle|\; (U_0,P_0,H_0)=(0,p,h)\right],\end{aligned} \tag{76}$$

which is the advantage of the risky arm over the safe arm. For each $u\in\mathbb{U}$, let $h_u$ be the deterministic value which $H_{\delta}$ attains after an initial choice of $u$. By Assumption 12, the inequality $h_0 \le h \le h_1$ holds. Furthermore, $P_{\delta} = P_0$ holds under an initial choice $u = 0$. By the monotonicity and convexity result of Lemma 6,

$$\begin{aligned}&\mathbb{E}\left[V^{\delta}(P_{\delta},H_{\delta}) \;\middle|\; (U_0,P_0,H_0)=(1,p,h)\right] - \mathbb{E}\left[V^{\delta}(P_{\delta},H_{\delta}) \;\middle|\; (U_0,P_0,H_0)=(0,p,h)\right]\\ &\quad= \mathbb{E}\left[V^{\delta}(P_{\delta},h_1) \;\middle|\; (U_0,P_0,H_0)=(1,p,h)\right] - V^{\delta}(p,h_0)\\ &\quad\ge \mathbb{E}\left[V^{\delta}(P_{\delta},h_1) \;\middle|\; (U_0,P_0,H_0)=(1,p,h)\right] - V^{\delta}(p,h_1) \ge 0.\end{aligned} \tag{77}$$

It follows that (76) is strictly positive if $\overline{b}^{\delta}(1,p,h) > \overline{b}^{\delta}(0,p,h)$. In this case, the initial choice of the risky arm is uniquely optimal. □

**Lemma 8** (Optimality of stopping rules)**.** *For each $\delta > 0$, $V^{\delta}(p,h)$ is a supremum over values of stopping rules.*

*Proof. Step 1 (Discrete setting).* We fix $\delta > 0$ and work on the uniform time grid $t_i = i\delta$, $i\in\mathbb{N}$. The one-stage payoff of the problem with partial observations is given by

$$b^{\delta}(u,x,h) = \int_0^{\delta} \rho e^{-\rho t} b(u,x,H_t)\mathrm{d}t, \tag{78}$$

where $H$ stems from a control with partial observations with constant control process $U_t \equiv u$ and initial condition $(X, H_0) = (x, h)$. The one-stage payoff of the safe arm is

$$k^\delta = b^\delta(0, x, h) = \int_0^\delta \rho e^{-\rho t} k \mathrm{d}t. \tag{79}$$

By abuse of notation, we identify indices $i \in \mathbb{N}$ with times $t_i$, writing $U_i$ for the value of $U$ on $(t_i, t_{i+1}]$ and $(P_i, H_i, R_i)$ for the value of $(P, H, R)$ at $t_i$.

*Step 2 (Finite horizon).* We truncate the problem with partial observations to a finite time horizon $n$. In the truncated problem, the value of a control $(U, X, H, R)$ is given by

$$J^{\text{p.o.}}(U, X, H, R) = \mathbb{E}\left[\sum_{i=0}^{n} e^{-\rho\delta i} b^\delta(U_i, X, H_i)\right]. \tag{80}$$

We will show by induction on $n$ that there exists an optimal stopping rule, i.e., a control that never switches from safe to the risky arm. For $n = 0$, there is nothing to prove. Now let $(U, X, H, R)$ be an optimal control for the problem with horizon $n + 1$ constructed via Lemma 3 from an optimal Markovian control for the truncated separated problem. As $H$ evolves deterministically given $U$, it is possible to write $U = F(R)$ for a piecewise constant process $F$ on the path space $D_{\mathbb{R}}[0, \infty)$.[18] The inductive hypothesis allows one to assume that for $i \geq 1$, $U_i$ never switches from the safe to the risky arm. If $U_0$ indicates the risky arm, the proof is complete. Otherwise, $U$ has the form

$$U_i = \begin{cases} 0, & \text{if } i = 0 \text{ or } i > T, \\ 1, & \text{if } 1 \leq i \leq T, \end{cases} \tag{81}$$

for some stopping time $T$. Given that the safe arm is chosen initially, the reward process $R$ is deterministic during the first stage. Therefore, there is a modification of $F$ that does not depend on the path of $R$ on the interval $[0, \delta]$. This makes it possible to define an adapted process $F^*$ which skips the first action of $F$. Then $F$ is a stopping rule. Formally, $F^*$ can be defined as

$$F^*(R) = \mathcal{S}^\delta F\big(\mathcal{S}^{-\delta} R\big), \tag{82}$$

where for any process $Z$, $(\mathcal{S}^\delta Z)_t = Z_{(t+\delta)\vee 0}$ is a shift of $Z$ by $\delta$. As the martingale problem $(\mathcal{A}, F^*)$ is well-posed, there is a corresponding control $(U^*, X^*, H^*, R^*)$ with $U^* = F^*(R^*)$ and initial condition $\mathbb{E}[X^*] = \mathbb{E}[X], H_0^* = H_0$. For comparison, we also define $(U^0, X^0, H^0, R^0)$ as the control where the risky arm $U^0 \equiv 0$ is chosen all the time, still with the same initial condition $\mathbb{E}[X^0] = \mathbb{E}[X], H_0^0 = H_0$. The values of the controls are denoted by $J, J^*$, and $J^0$, respectively.

[18] This is easily seen for $U_0$, which is deterministic. For $U_{i+1}$, it follows by induction because $H_{i+1}$ is a deterministic function of $U_i$.

Then

$$J^* - J^0 = \mathbb{E}\left(\sum_{i=0}^{T-1} e^{-\rho\delta i}\big(b^\delta(1,X,H_i^*) - k^\delta\big)\right) \geq \mathbb{E}\left(\sum_{i=1}^{T} e^{-\rho\delta(i-1)}\big(b^\delta(1,X,H_i) - k^\delta\big)\right), \tag{83}$$

$$J - J^0 = \mathbb{E}\left(\sum_{i=1}^{T} e^{-\rho\delta i}\big(b^\delta(1,X,H_i) - k^\delta\big)\right) \geq 0. \tag{84}$$

The first inequality holds because choosing the safe arm decreases $H$, see Assumption 12. The second inequality holds because the value $J$ of the optimal control is at least as high as $J^0$. Thus,

$$\begin{aligned} J^* - J \geq\ & \mathbb{E}\left[\sum_{i=1}^{T}\left(e^{-\rho\delta(i-1)} - e^{-\rho\delta i}\right)\left(b^\delta(1,X,H_i) - k^\delta\right)\right] \\ &= \sum_{i=1}^{\infty}\left(e^{-\rho\delta(i-1)} - e^{-\rho\delta i}\right)\underbrace{\mathbb{E}\Big[\mathbb{1}_{i\leq T}\big(b^\delta(1,X,H_i) - k^\delta\big)\Big]}_{=:b_i}. \end{aligned} \tag{85}$$

The increments of $(b_i)_{i\in\mathbb{N}}$ are given by

$$b_{i+1} - b_i = \mathbb{E}\Big[\mathbb{1}_{i\leq T}\big(b^\delta(1,X,H_{i+1}) - b^\delta(1,X,H_i)\big)\Big] + \mathbb{E}\Big[\mathbb{1}_{i=T}\big(k^\delta - b^\delta(1,X,H_{i+1})\big)\Big]. \tag{86}$$

The first summand on the right-hand side is non-negative for $i \geq 1$ because $H$ increases while the risky arm is played. By the $\mathcal{F}_{i+1}^R$-measurability of $\mathbb{1}_{i=T}$ and $H_{i+1}$, the second summand can be written as

$$\mathbb{E}\Big[\mathbb{1}_{i=T}\big(k^\delta - b^\delta(1,X,H_{i+1})\big)\Big] = \mathbb{E}\Big[\mathbb{1}_{i=T}\big(k^\delta - \overline{b}^\delta(1,P_{i+1},H_{i+1})\big)\Big] = \mathbb{E}\Big[\mathbb{1}_{i=T}\big(k^\delta - \overline{b}^\delta(1,P_{T+1},H_{T+1})\big)\Big], \tag{87}$$

where $\overline{b}^\delta(u,p,h)$ is defined in Equation (74). As it is optimal under $U$ (see Equation (81)) to choose the safe arm at stage $T+1$, the inequality $k^\delta \geq \overline{b}^\delta(1,P_{T+1},H_{T+1})$ holds by Lemma 7. This proves $b_{i+1} \geq b_i$, for all $i \geq 1$. By Equation (84), we also have

$$\sum_{i=1}^{\infty} e^{-\rho\delta i} b_i \geq 0. \tag{88}$$

By Berry and Fristedt [7, Equation (5.2.8)] this implies

$$J^* - J = \sum_{i=1}^{\infty}\left(e^{-\rho\delta(i-1)} - e^{-\rho\delta i}\right) b_i \geq 0, \tag{89}$$

since truncated geometric discount sequences are regular. Thus we have constructed an optimal stopping rule $(U^*, X^*, H^*, R^*)$ for the truncated problem with horizon $n+1$.

*Step 2 (Infinite horizon).* We have shown that stopping rules are optimal for each discretized problem with finite horizon $n$. It follows by approximation that the value function $V^\delta(p,h)$ of the

discretized problem with infinite horizon is a supremum over stopping rules. The argument can be found in the proof of Berry and Fristedt [7, Theorem 5.2.2]. □

**Lemma 9** (Description of optimal stopping rules)**.** *The stopping time* $T^* = \inf\{t : V(P_t, H_t) \le k\}$ *is optimal for the separated problem.*

*Proof.* For each $(p, h) \in [0, 1] \times \mathbb{H}$, there is a unique solution $(P, H, R)$ of the martingale problem for $(\mathcal{G}, 1)$ by Assumption 9. The family $(P, H)$ of processes, indexed by the initial condition $(p, h)$, is a Feller process. This follows from Jacod and Shiryaev [26, Theorem IX.4.39] using similar arguments as in Step 3 of the proof of Lemma 2. Let $(\widetilde{P}, \widetilde{H})$ be the killed version of $(P, H)$ with killing rate $\rho$ and let $\Delta$ denote the "cemetery point" of the killed process. We refer to Peskir and Shiryaev [49, Section II.5.4] for the terminology. Let $\overline{b}(u, \Delta) = 0$ and

$$A_t = A_0 + \int_0^t \big(\overline{b}(1, \widetilde{P}_t, \widetilde{H}_t) - k\big)\mathrm{d}t. \tag{90}$$

Then $Z = (\widetilde{P}, \widetilde{H}, A)$ is a Feller process on the state space $\mathbb{Z} = ([0, 1] \times \mathbb{H} \cup \{\partial\}) \times \mathbb{R}$. Let $(\mathbb{P}_z)_{z \in \mathbb{Z}}$ denote the family of laws of $Z$ starting from the initial condition $Z_0 = z$. There is an associated family of stopping problems

$$W(z) = \sup_T \mathbb{E}_z(A_T), \tag{91}$$

where the supremum is taken over all $\{\mathcal{F}_t^Z\}$-stopping times. For any $z = (p, h, a) \neq \Delta$,

$$\begin{aligned} W(z) &= \sup_T \mathbb{E}_{(p,h,a)}[A_T] = \sup_T \mathbb{E}_{(p,h,0)}[A_T] + a \\ &= \sup_T \mathbb{E}_{(p,h,0)} \left[ \int_0^T \rho e^{-\rho t} \big(\overline{b}(1, P_t, H_t) - k\big)\mathrm{d}t \right] + a = V(p, h) - k + a, \end{aligned} \tag{92}$$

because $V(p, y)$ is a supremum of values of stopping rules by part (a) of Theorem 2. The stopping set $\mathbb{D} \subset \mathbb{Z}$ is defined as in Peskir and Shiryaev [49, Equation (2.2.5)] by

$$\mathbb{D} = \big\{z = (p, h, a) \in \mathbb{Z} \colon W(z) \le a\big\} = \Big(\big\{(p, h) \in [0, 1] \times \mathbb{H} \colon V(p, h) \le k\big\} \cup \{\Delta\}\Big) \times \mathbb{R}. \tag{93}$$

The last equality holds because $W(\partial, a) = a$ by definition. The function $W$ is lower semi-continuous by Peskir and Shiryaev [49, Equation (2.2.80)] because $(\widetilde{P}, \widetilde{H}, A)$ is Feller. Therefore, the set $\mathbb{D}$ is closed. Then the right-continuity of the filtration implies that

$$T^* = \inf\{t \ge 0 : X_t \in D\} = \inf\{t : V(P_t, H_t) \le k\} \tag{94}$$

is a stopping time. Note that $\Delta \in D$, which implies $\mathbb{P}(T^* < \infty) = 1$. Then Peskir and Shiryaev [49, Corollary 2.9] implies that $T^*$ is optimal. □

**Lemma 10** (Asymptotic learning)**.** *Assume* $0 < P_0 < 1$*. Then the following statements hold for any control* $(U, X, H, R)$ *of the problem with partial observations and the corresponding belief*

*process $P$.*

*(a) Assume that the measures $K_R(1,h,\cdot)$ and $K_R(0,h,\cdot)$ are equivalent for all $h$. Then learning in finite time is impossible, i.e., $0 < P_t < 1$ holds a.s. for all $t \geq 0$. Moreover, asymptotic learning does not occur if the agent invests only a finite amount into the risky arm, i.e.,*

$$\{\textstyle\int_0^\infty U_t dt < \infty\} \subseteq \{0 < P_\infty < 1\} \quad \mathbb{P}\text{-a.s.} \tag{95}$$

*(b) Assume that $\Phi(1,\cdot)$ is bounded from below by a positive constant. Then asymptotic learning is guaranteed if the agent invests an infinite amount in the risky arm, i.e.,*

$$\{\textstyle\int_0^\infty U_t dt = \infty\} \subseteq \{P_\infty = X\} \quad \mathbb{P}\text{-a.s.} \tag{96}$$

*(c) If the conditions of (a) and (b) are satisfied, then asymptotic learning occurs if and only if the agent invests an infinite amount in the risky arm:*

$$\{\textstyle\int_0^\infty U_t dt = \infty\} = \{P_\infty = X\} \quad \mathbb{P}\text{-a.s.} \tag{97}$$

*Proof. Step 1 (Hellinger process).* Let $\mathbb{P}_1$ and $\mathbb{P}_0$ be defined by conditioning the measure $\mathbb{P}$ on the events $X = 1$ and $X = 0$, respectively. We want to calculate the Hellinger process $h(\frac{1}{2})$ of order $\frac{1}{2}$ of the measures $\mathbb{P}_1$ and $\mathbb{P}_0$. Let $P_t = \mathbb{E}[X \mid \mathcal{F}_t^Y]$ be the belief process. By Equation (7), $P/P_0$ is the density process of $\mathbb{P}_1$ relative to $\mathbb{P}$. Similarly, $(1-P)/(1-P_0)$ is the density process of $\mathbb{P}_0$ relative to $\mathbb{P}$. For all $p, q \in \mathbb{R}$, let

$$\psi(p,q) = \frac{p+q}{2} - \sqrt{pq} \tag{98}$$

and let $\nu(\mathrm{d}t, \mathrm{d}p, \mathrm{d}q)$ be the compensator of the integer-valued random measure associated the jumps of $(P, 1-P)$. Let $S$ be the first time that $P$ or $P_-$ hits zero or one,

$$S = \inf\big\{t \geq 0 : P_t \in \{0,1\} \text{ or } P_{t-} \in \{0,1\}\big\}. \tag{99}$$

By Jacod and Shiryaev [26, Lemma III.3.7], $P$ is constant on $[\![S, \infty[\![$. Therefore, on this interval, $\langle P^c, P^c\rangle$ is constant and $\nu$ has no charge. After canceling out the terms $P_0$ and $(1-P_0)$, the formula

for $h(\frac{1}{2})$ given in Jacod and Shiryaev [26, Theorem IV.1.33] reads as

$$\begin{aligned}
h(\tfrac{1}{2}) &= \frac{1}{8}\left(\frac{1}{P_-^2}\bullet\langle P^c,P^c\rangle - \frac{2}{P_-(1-P_-)}\bullet\langle P^c,1-P^c\rangle + \frac{1}{(1-P_-)^2}\bullet\langle 1-P^c,1-P^c\rangle\right)\\
&\quad + \psi\left(1+\frac{p}{P_-},1+\frac{q}{1-P_-}\right)*\nu(\mathrm{d}t,\mathrm{d}p,\mathrm{d}q)\\
&= \frac{1}{8}\left(\frac{1}{P_-}+\frac{1}{1-P_-}\right)\bullet\langle P^c,P^c\rangle\\
&\quad + \psi\left(1+\frac{j(U,P_-,Y_-,z)}{P_-},1-\frac{j(U,P_-,Y_-,z)}{1-P_-}\right)*\overline{K}(U,P_-,Y_-,\mathrm{d}z)\mathrm{d}t\\
&= \frac{1}{8}\phi_1(U,Y)^\top\sigma^2(U,Y)\phi_1(U,Y)\bullet I^S\\
&\quad + \psi\Bigg(\frac{\phi_2(U,Y_-,z)}{P_-\phi_2(U,Y_-,z)+(1-P_-)\big(2-\phi_2(U,Y_-,z)\big)},\\
&\qquad\qquad \frac{2-\phi_2(U,Y_-,z)}{P_-\phi_2(U,Y_-,z)+(1-P_-)\big(2-\phi_2(U,Y_-,z)\big)}\Bigg)\mathbb{1}_{[\![0,S]\!]}*\overline{K}(U,P_-,Y_-,\mathrm{d}z)\mathrm{d}t\\
&= \frac{1}{8}\phi_1(U,Y)^\top\sigma^2(U,Y)\phi_1(U,Y)\bullet I^S\\
&\quad + \int\frac{1-\sqrt{\phi_2(U,Y,z)\big(2-\phi_2(U,Y,z)\big)}}{P\phi_2(U,Y,z)+(1-P)\big(2-\phi_2(U,Y,z)\big)}\overline{K}(U,P,Y,\mathrm{d}z)\bullet I^S\\
&= \frac{1}{8}\phi_1(U,Y)^\top\sigma^2(U,Y)\phi_1(U,Y)\bullet I^S\\
&\quad + \int\left(1-\sqrt{\phi_2(U,Y,z)\big(2-\phi_2(U,Y,z)\big)}\right)\overline{K}(U,1/2,Y,\mathrm{d}z)\bullet I^S = \Phi(U,Y)\bullet I^S,
\end{aligned}\tag{100}$$

where $\Phi$ is defined in Assumption 5.

*Step 2 (Finite investment prevents asymptotic learning).* We define stopping times $T$ and $T_n$ as in Equations (34) and (15). $T$ is the first time that $P$ or $P_-$ hits zero and $T_n$ announces $T$. Let us assume for contradiction that $P$ jumps to zero, i.e., $P_{T-}>0$. Then $T_n=T$ holds for all sufficiently large $n$. Consequently, the process $D^n=\mathcal{E}(L^n)=P^{T_n}/P_0$ defined in Equation (17) also jumps to zero. Therefore, $L^n$ has a jump of height $-1$. This is not possible because $\phi_2(u,y,z)>0$ holds by the assumption that $K(u,1,y,\cdot)$ and $K(u,0,y,\cdot)$ are equivalent. This proves that $P$ does not jump to zero. A similar argument where the rôles of $\mathbb{P}_0$ and $\mathbb{P}_1$ are reversed shows that $P$ cannot jump to one. It follows that for any stopping time $\tau$,

$$\{h(\tfrac{1}{2})_\tau=\infty\}=\{S\le\tau,P_{S-}=0\}=\{P_\tau=0\}=\{P_\tau=0 \text{ or } P_\tau=1\}\quad \mathbb{P}_0\text{-a.s.},\tag{101}$$

$$\{h(\tfrac{1}{2})_\tau=\infty\}=\{S\le\tau,P_{S-}=1\}=\{P_\tau=1\}=\{P_\tau=0 \text{ or } P_\tau=1\}\quad \mathbb{P}_1\text{-a.s.}\tag{102}$$

In Equations (101) and (102), the first equality holds by Schachermayer and Schachinger [57, Theorem 1.5]. This theorem states that the divergence of the Hellinger process is equivalent to the mutual singularity of the measures $\mathbb{P}_1$ and $\mathbb{P}_0$, but in such a way that the singularity is not

obtained by a sudden jump of the density process to zero or one. The second equality holds because such jumps are not possible by the previous claim. For the third equality, see Jacod and Shiryaev [26, Proposition III.3.5.(ii)]. By Assumption 10, the safe arm reveals no information about the hidden state $X$, resulting in $\Phi(0, y) = 0$. Together with Assumption 5 bounding $\Phi$ from above, Equations (101) and (102) imply

$$\{\textstyle\int_0^\infty U_t dt < \infty\} \subseteq \{h(\frac{1}{2})_\infty < \infty\} = \{0 < P_\infty < 1\} \quad \mathbb{P}\text{-a.s.} \tag{103}$$

This proves (a).

*Step 3 (Infinite investment induces asymptotic learning).* Let $\tau$ be a stopping time. If $S$ does not occur before $\tau$ and $\int_0^\tau U_t dt = \infty$, then $h(\frac{1}{2})_\tau = \infty$ because of the lower bound $\inf_y \Phi(1, y) > 0$. Therefore,

$$\{\textstyle\int_0^\tau U_t dt = \infty\} \subseteq \{h(\frac{1}{2})_\tau < \infty\} \cup \{S \leq \tau\}. \tag{104}$$

Moreover, it follows from Schachermayer and Schachinger [57, Theorem 1.5] that

$$\{h(\tfrac{1}{2})_\tau < \infty\} \cup \{S \leq \tau\} = \{S \leq \tau, P_{S-} = 0\} \cup \{S \leq \tau\} = \{P_\tau = X\} \quad \mathbb{P}_0\text{-a.s.}, \tag{105}$$

$$\{h(\tfrac{1}{2})_\tau < \infty\} \cup \{S \leq \tau\} = \{S \leq \tau, P_{S-} = 1\} \cup \{S \leq \tau\} = \{P_\tau = X\} \quad \mathbb{P}_1\text{-a.s.} \tag{106}$$

It follows that

$$\{\textstyle\int_0^\tau U_t dt = \infty\} \subseteq \{P_\tau = X\} \quad \mathbb{P}\text{-a.s.}, \tag{107}$$

which proves (b). Finally, (c) follows from (a) and (b). □

# References


[1] Daron Acemoglu et al. "Bayesian Learning in Social Networks". In: *The Review of Economic Studies* 78.4 (Oct. 2011), p. 1201.

[2] Peter Auer et al. "The nonstochastic multiarmed bandit problem". In: *SIAM J. Comput.* 32.1 (2002/03), 48–77 (electronic).

[3] Jeffrey S. Banks and Rangarajan K. Sundaram. "Denumerable-Armed Bandits". In: *Econometrica* 60.5 (Sept. 1992), pp. 1071–1096.

[4] Jeffrey S. Banks and Rangarajan K. Sundaram. "Switching Costs and the Gittins Index". In: *Econometrica* 62.3 (May 1994), pp. 687–694.

[5] A. Basu, A. Bose, and JK Ghosh. *An Expository Review of Sequential Design and Allocation Rules*. Tech. rep. Department of Statistics, Purdue University, 1990.

[6] Dirk Bergemann and Juuso Välimäki. *Bandit Problems*. Cowles Foundation Discussion Papers 1551. Cowles Foundation for Research in Economics, Yale University, Jan. 2006.

[7] Donald A. Berry and Bert Fristedt. *Bandit problems: sequential allocation of experiments*. Monographs on statistics and applied probability. London; New York: Chapman and Hall, 1985.

[8] Patrick Bolton and Christopher Harris. "Strategic experimentation". In: *Econometrica* 67.2 (1999), pp. 349–374.

[9] Patrick Bolton and Christopher Harris. "Strategic Experimentation: The Undiscounted Case". In: *Incentives, Organization and Public Economics. Papers in Honour of Sir James Mirrlees*. Ed. by Peter J. Hammond and Gareth D. Myles. Oxford and New York: Oxford University Press, 2000, pp. 53–68.

[10] A. Cohen and E. Solan. "Bandit problems with Levy payoff processes". In: *Mathematics of Operations Research* 38.1 (Feb. 2013), pp. 92–107.

[11] Flavio Cunha and James J. Heckman. *Investing in Our Young People*. Working Paper 16201. National Bureau of Economic Research, July 2010. URL: `http://www.nber.org/papers/w16201`.

[12] Flavio Cunha et al. "Interpreting the evidence on life cycle skill formation". In: *Handbook of the Economics of Education* 1 (2006), pp. 697–812.

[13] Oliver Delzeith. "On Skorohod spaces as universal sample path spaces". In: *arXiv preprint math/0412092* (2004).

[14] Will Dobbie and Roland Fryer. *Getting beneath the veil of effective schools: Evidence from New York City*. Tech. rep. National Bureau of Economic Research, 2011.

[15] N. El Karoui and I. Karatzas. "Dynamic allocation problems in continuous time". In: *Ann. Appl. Probab.* 4.2 (1994), pp. 255–286.

[16] Nicole El Karoui, D. H. Nguyen, and Monique Jeanblanc-Picqué. "Existence of an optimal Markovian filter for the control under partial observations". In: *SIAM J. Control Optim.* 26.5 (1988), pp. 1025–1061.

[17] Y. Faihe and J.P. Müller. "Behaviors coordination using restless bandits allocation indexes". In: *From Animals to Animats 5 (Proc. 5th Int. Conf. Simulation of Adaptive Behavior)*. 1998, pp. 159–164.

[18] Wendell H. Fleming and Makiko Nisio. "On the existence of optimal stochastic controls". In: *J. Math. Mech.* 15 (1966), pp. 777–794.

[19] Wendell H. Fleming and Étienne Pardoux. "Optimal control for partially observed diffusions". In: *SIAM J. Control Optim.* 20.2 (1982), pp. 261–285.

[20] J.C. Gittins. "Bandit processes and dynamic allocation indices". In: *Journal of the Royal Statistical Society. Series B (Methodological)* (1979), pp. 148–177.

[21] John Gittins, Kevin Glazebrook, and Richard Weber. *Multi-armed bandit allocation Indices*. Wiley-Blackwell, 2011.

[22] KD Glazebrook, C. Kirkbride, and D. Ruiz-Hernandez. "Spinning plates and squad systems: policies for bi-directional restless bandits". In: *Advances in applied probability* 38.1 (2006), pp. 95–115.

[23] KD Glazebrook, J. Nino-Mora, and PS Ansell. "Index policies for a class of discounted restless bandits". In: *Advances in Applied Probability* 34.4 (2002), pp. 754–774.

[24] KD Glazebrook, D. Ruiz-Hernandez, and C. Kirkbride. "Some indexable families of restless bandit problems". In: *Advances in Applied Probability* 38.3 (2006), pp. 643–672.

[25] Jean Jacod and Jean Mémin. "Sur un type de convergence intermédiaire entre la convergence en loi et la convergence en probabilité". In: *Séminaire de Probabilités XV 1979/80*. Springer, 1981, pp. 529–546.

[26] Jean Jacod and Albert N. Shiryaev. *Limit theorems for stochastic processes*. Second. Vol. 288. Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]. Berlin: Springer-Verlag, 2003.

[27] Tackseung Jun. "A survey on the bandit problem with switching costs". In: *De Economist* 152 (4 2004), pp. 513–541.

[28] I. Karatzas. "Gittins indices in the dynamic allocation problem for diffusion processes". In: *The Annals of Probability* (1984), pp. 173–192.

[29] Godfrey Keller and Sven Rady. "Breakdowns". In: *Theoretical Economics* 10.1 (2015), pp. 175–202.

[30] Godfrey Keller and Sven Rady. "Strategic experimentation with Poisson bandits". In: *Theoretical Economics* 5.2 (May 2010), pp. 275–311.

[31] Godfrey Keller, Sven Rady, and Martin Cripps. "Strategic experimentation with exponential bandits". In: *Econometrica* 73.1 (2005), pp. 39–68.

[32] GP Klimov. "Time-sharing service systems. I". In: *Theory of Probability & Its Applications* 19.3 (1975), pp. 532–551.

[33] M. Kohlmann. "Existence of optimal controls for a partially observed semimartingale". In: *Stochastic Processes and their Applications* 13.2 (1982), pp. 215–226.

[34] T.G. Kurtz and D.L. Ocone. "Unique characterization of conditional distributions in nonlinear filtering". In: *The Annals of Probability* (1988), pp. 80–107.

[35] Thomas G. Kurtz. "Martingale problems for conditional distributions of Markov processes". In: *Electron. J. Probab.* 3.9 (1998), pp. 1–29.

[36] Thomas G. Kurtz. "Martingale problems for controlled processes". In: *Stochastic modelling and filtering*. Springer, 1987, pp. 75–90.

[37] Thomas G Kurtz and Giovanna Nappo. "The filtered martingale problem". In: *The Oxford Handbook of Nonlinear Filtering* (2011), pp. 129–168.

[38] H.J. Kushner and P.G. Dupuis. *Numerical methods for stochastic control problems in continuous time*. Vol. 24. Springer, 2000.

[39] BF La Scala and B. Moran. "Optimal target tracking with restless bandits". In: *Digital Signal Processing* 16.5 (2006), pp. 479–487.

[40] Dominique Lépingle and Jean Mémin. "Sur l'intégrabilité uniforme des martingales exponentielles". In: *Z. Wahrsch. Verw. Gebiete* 42.3 (1978), pp. 175–203.

[41] Chih-ping Li and Michael J Neely. "Network utility maximization over partially observable markovian channels". In: *Performance Evaluation* (2012).

[42] A. Mahajan and D. Teneketzis. "Multi-armed bandit problems". In: *Foundations and Applications of Sensor Management* (2008), pp. 121–151.

[43] Avi Mandelbaum. "Continuous multi-armed bandits and multiparameter processes". In: *Ann. Probab.* 15.4 (1987), pp. 1527–1556.

[44] B. P. McCall and J. J. McCall. "A Sequential Study of Migration and Job Search". In: *Journal of Labor Economics* 5.4 (1987), pp. 452–476.

[45] Hiroaki Morimoto. "On average cost stopping time problems". In: *Probab. Theory Related Fields* 90.4 (1991), pp. 469–490.

[46] J. Nino-Mora. "Restless bandits, partial conservation laws and indexability". In: *Advances in Applied Probability* 33.1 (2001), pp. 76–98.

[47] Jerome Le Ny, Munther Dahleh, and Eric Feron. "Multi-UAV dynamic routing with partial observations using restless bandit allocation indices". In: *American Control Conference, 2008*. IEEE. 2008, pp. 4220–4225.

[48] Sandeep Pandey, Deepayan Chakrabarti, and Deepak Agarwal. "Multi-armed bandit problems with dependent arms". In: *Proceedings of the 24th international conference on Machine learning*. ACM. 2007, pp. 721–728.

[49] Goran Peskir and Albert Shiryaev. *Optimal stopping and free-boundary problems*. Lectures in Mathematics ETH Zürich. Birkhäuser Verlag, 2006.

[50] Huyên Pham. "Optimal stopping of controlled jump diffusion processes: a viscosity solution approach". In: *J. Math. Systems Estim. Control* 8.1 (1998).

[51] Huyên Pham. "Optimal stopping of controlled jump diffusion processes and viscosity solutions". In: *C. R. Acad. Sci. Paris Sér. I Math.* 320.9 (1995), pp. 1113–1118.

[52] W.B. Powell. *Approximate Dynamic Programming: Solving the curses of dimensionality*. Vol. 703. Wiley-Interscience, 2007.

[53] È. L. Presman and I. N. Sonin. *Sequential control with incomplete information*. Economic Theory, Econometrics, and Mathematical Economics. San Diego, CA: Academic Press Inc., 1990.

[54] Ernst L Presman. "Poisson version of the two-armed bandit problem with discounting". In: *Theory of Probability & Its Applications* 35.2 (1990), pp. 307–317.

[55] Herbert Robbins. "Some aspects of the sequential design of experiments." In: *Bull. Am. Math. Soc.* 58 (1952), pp. 527–535.

[56] Michael Rothschild. "A two-armed bandit theory of market pricing". In: *J. Econom. Theory* 9.2 (1974), pp. 185–202.

[57] W. Schachermayer and W. Schachinger. "Is there a predictable criterion for mutual singularity of two probability measures on a filtered space?" In: *Teor. Veroyatnost. i Primenen.* 44.1 (1999), pp. 101–110.

[58] Atle Seierstad. *Stochastic control in discrete and continuous time*. New York: Springer, 2009.

[59] Mariusz Startek. "Vague Convergence in the Skorohod Representation Theorem". In: *Int. J. Contemp. Math. Sciences* 7.22 (2012), pp. 1061–1066.

[60] Richard H. Stockbridge. "A separation principle for partially observed control of singular stochastic processes". In: *Nonlinear Analysis* 63 (2005), e2057–e2065.

[61] Daniel W. Stroock and S. R. Srinivasa Varadhan. *Multidimensional diffusion processes*. Classics in Mathematics. Reprint of the 1997 edition. Berlin: Springer-Verlag, 2006, pp. xii+338.

[62] Michael H Veatch and Lawrence M Wein. "Scheduling a make-to-stock queue: Index policies and hedging points". In: *Operations Research* 44.4 (1996), pp. 634–647.

[63] R. Washburn. "Application of multi-armed bandits to sensor management". In: *Foundations and Applications of Sensor Management* (2008), pp. 153–175.

[64] R.R. Weber and G. Weiss. "Addendum to 'On an index policy for restless bandits'". In: *Advances in Applied probability* (1991), pp. 429–430.

[65] R.R. Weber and G. Weiss. "On an index policy for restless bandits". In: *Journal of Applied Probability* (1990), pp. 637–648.

[66] Martin L. Weitzman. "Optimal Search for the Best Alternative". In: *Econometrica* 47.3 (1979), pp. 641–654.

[67] P. Whittle. "Arm-Acquiring Bandits". In: *The Annals of Probability* 9.2 (1981), pp. 284–292.

[68] P. Whittle. "Restless bandits: activity allocation in a changing world". In: *J. Appl. Probab.* 25 (1988), pp. 287–298.

[69] W. M. Wonham. "On the separation theorem of stochastic control". In: *SIAM J. Control* 6 (1968), pp. 312–326.