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

Roland G. Fryer, Jr.
Harvard University and NBER

Philipp Harms
Education Innovation Laboratory

February 2013

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 and job search. 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, Lawrence Katz, Peter Michor, Derek Neal, Yuliy Sannikov, Andrei Shleifer, Mete Soner, and seminar participants at Barcelona GSE and Harvard University for helpful comments. Brad Allan, Matthew Davis, Alexander Dubbs, Blake Heller, Sue Lin, and Madhav Mani provided excellent research assistance. 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@edlabs.harvard.edu [Harms]. The usual caveat applies.

1 Introduction

Bandit models are sequential decision problems where, at each stage, a resource like time, effort or money has to be allocated strategically between several options, typically 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 math 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 1974; Berry and Fristedt 1985; Bolton and Harris 1999; Keller and Rady 2010). To understand how firms set prices without a clear understanding of their demand curves, Rothschild (1974) 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 (1985) 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 (1999, 2000). 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 e.a. (2005, 2010).

An important assumption in the classical bandit literature is that the reward distributions of arms that are not chosen do not evolve; they rest (Gittins, Glazebrook, and Weber 2011). 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 (2006) 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.

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, which typically increases her chances of successful future job applications. Conversely, not actively

1. The importance of relaxing this assumption has been recognized early on in the seminal work of Whittle (1988), who proposed clinical trials, aircraft surveillance and assignment of workers to tasks as potential applications.

2. Cunha et al. (2006) make a similar argument in a different context.

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] However, when a certain indexability condition is met, Whittle's (1988) index can lead to approximately optimal solutions (Weber and Weiss 1990, 1991). This index plays the same fundamental role for restless bandits that Gittins' (1979) index has for classical ones: it decomposes the task of solving multi-armed bandits into multiple tasks of solving bandits with one known and one unknown arm. The known arm yields constant rewards and can be interpreted as a cost of investment in the unknown 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 (1999), Keller and Rady (2010), and Cohen and Solan (2013) that we extend to the restless case. In these pioneering works, the reward from the unknown arm is Brownian motion, a Poisson process, or a Levy process. The unobserved quantity is a Bernoulli variable called the "type" of the agent. Optimal strategies are found by constructing explicit solutions of the Hamilton-Jacobi-Bellman (HJB) equation, which is a non-linear second-order differential-difference equation. 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 unknown 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 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 to this equation can be found.

Using new insights from stochastic control theory, along with a monotonicity condition on the restless arm (e.g. today's investments make tomorrow's investments more profitable, while a lack of investment today decreases the profits of future investments), this paper establishes two results. The first result is a *separation theorem* (theorem 1) that establishes the equivalence of the optimal control problem with partial observations to a fully observed control problem called the separated control problem. The fact that this equivalence holds is crucial for the solution of the problem and is used in many works, including Bolton and Harris (1999); Keller and Rady (2010); Cohen and

3. Bandits where the active and passive action have opposite effects on payoffs are called *bi-directional bandits* (Glazebrook, Kirkbride, and Ruiz-Hernandez 2006), 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 2000; Powell 2007; Nino-Mora 2001).

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

Solan (2013). Standard formulations of the partially observable control problem involving Zakai's transform (Fleming and Pardoux 1982) or time changes (El Karoui and Karatzas 1994) do not work for restless bandit problems. However, we show that the frameworks of Fleming and Nisio (1966); Wonham (1968); Kohlmann (1982) can be used and extended to general semimartingale observation processes. We describe these issues in detail since they are rarely discussed in the context of bandit problems. In our second, and main, result (theorem 2), we establish the *optimality of stopping rules* and characterize optimal strategies by an index that formally coincides with Gittins' index. This allows us to deduce comparative statics which imply the indexability of our model in the sense of Whittle (1988). We provide a sketch of our approach below.

The first sections of the analysis are dedicated to a rigorous setup of the stochastic control problem in continuous time (sections 3.3-3.5). The bandit problem is first formulated as a problem of stochastic optimal control under *partial observations*. The structure of our model requires a strong measurability condition on controls to ensure that they do not depend on the unobserved type, see section 3.3. The condition has been used in the early literature on optimal control (Fleming and Nisio 1966; Wonham 1968; Kohlmann 1982) but came out of favor because it entails difficulties in establishing the existence of optimal controls.

We are able to circumvent these difficulties by showing that regardless of the strong measurability constraint, the partially observable problem is equivalent to a so-called *separated problem* with full observations. In the separated problem, admissibility of controls can be defined as usual and the existence of optimal controls is well-known. 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 the high type. The equivalence of the partially observable and the separated problem is established in theorem 1. Notice: the monotonicity condition is not needed there.

Our second, and main, result is that *stopping rules* are optimal. This result hinges on the monotonicity condition and is established in theorem 2. Our proof is based on a direct investigation of the sample paths of optimal strategies and an evaluation of the benefits of investing in the unknown arm sooner rather than later. This interchange argument was originally developed by Berry and Fristedt (1985) for classical bandit models, but the monotonicity assumption on the payoffs is precisely what is needed to make the argument work in the more general setting of restless bandits.

The result on optimal stopping means that it always better to invest first in the unknown arm and then in the known arm instead of the other way round. Intuitively, this sequence of investments matters for two reasons. First, investments in the unknown arm reveal information about the distribution of rewards. The sooner this information becomes available, the better. Second, early investments in the known arm deteriorate the rewards of later investments in the unknown arm. By contrast, early investments in the unknown arm do not make the known arm any less profitable. It follows that agents find it optimal to invest in the unknown arm initially.

They then switch to the known arm and never start investing in the unknown arm again.[6] The time where this change occurs is a stopping time that depends on the history of obtained rewards.

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 (1979) on classical bandits, but unused 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 known arm enlarge the set of states where the known arm is optimal, which means that our bandit model is *indexable* in the sense of Whittle (1988). 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]

To explain the structure of optimal strategies, we consider how information is processed by the 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 unknown 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 unknown or known arm. Namely, there is a curve in the belief–human capital domain such that it is optimal to invest in the unknown 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 known arm. This follows from the index representation of optimal strategies. The curve is called the *decision frontier*. It can be characterized as a level set of the index or value function.

Similar to classical bandit model, the dynamics of belief and human capital depend on the position relative to the decision frontier. There is an important, and potentially empirically relevant, difference: below the frontier, agents do not obtain any new information, and their belief remains constant, while their human capital *decreases* continually. In other words, 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: (1) they are initiated when students get close to the decision frontier, or (2) force inframarginal students to invest in the unknown arm (e.g. some charter schools (Dobbie and Fryer 2011)). Consistent with Cunha et al. (2006), 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

6. This argument is made rigorous in the proof of theorem 2.

7. Some sensor management models are indexable and have a continuous state space after their transformation to fully observed Markov decision problems (Washburn 2008). However, this is not the case in their formulation as partially observable control problems.

time where they hit the frontier. This happens when they encounter a series of bad outcomes from the unknown 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 unknown arm forever and learn their true type in the limit. In fact, under reasonable assumptions, investing in the unknown arm for an infinite amount of time is necessary and sufficient for *asymptotic learning* to occur (see theorem 3). To summarize, agents of the low type eventually end up choosing the known arm, which is optimal for them. However, high type agents can get discouraged by bad luck and stop investing in the unknown arm, even though the unknown arm would be optimal. Possible limit points of agents' trajectories in the belief–human capital space are depicted in Figures 1 and 2.

Our paper makes four contributions to the existing literature. First, we present an extension of classical bandit models of investment under uncertainty motivated by dynamic aspects of resource development. The model is new and has economic significance in a wide range of real world settings. As an example, we present how our model can be used to describe the economics of investment in education and discuss some policy potential implications. Second, we discover 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 2006), do not allow history-dependent payoffs (Washburn 2008), or work with very specific reward processes (e.g. Markov chains on finite state spaces as in Nino-Mora (2001)). Third, we deal with the delicate issue of setting up the partially observable control problem in continuous time. Recent standard formulations of optimal control under partial observation do not apply in our setting. We rediscover a framework that has been used mostly in the early control literature and show that it meets the needs of both classical and restless bandit models. In addition to its importance to the theory of optimal control, this is also a contribution to the bandit literature. Fourth, we present an unconventional approach to solve the bandit model. While the work horse of most of the bandit literature is either the Hamilton-Jacobi-Bellman equation or a setup using time changes, our argument is based on an investigation of the sample paths of optimal strategies. We resort to this approach because the other two approaches are not well adapted to the generality of our model, in particular the new dynamics.

The paper is structured as follows. Section 2 provides a brief review of the bandit literature in economics and applied math. Section 3.1 provides the model and section 3.2 connects our formulation, using semimartingales, to classical bandit models. A precise formulation of the partially observable problem is developed in section 3.3. The separated problem is defined in section 3.4 and the equivalence of the two problems is established in section 3.5. In section 3.6, it is shown that both problems are equivalent to optimal stopping. Furthermore, optimal strategies are characterized in terms of Gittins' index. In sections 3.7 and 3.8, asymptotic learning and long term limits of the belief process are studied. Finally, section 4 concludes.

2 A brief review of the multi-armed bandit literature

2.1 Models

Originally developed by Robbins (1952), 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 (1974), 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 1987; Weitzman 1979; Berry and Fristedt 1985; Li and Neely 2011; Banks and Sundaram 1992).

Classical bandits with reward processes driven by Brownian motion or a Poisson process were first solved by Karatzas (1984) and Presman (1990). Subsequently, Bolton and Harris (1999, 2000) and Keller e.a. (2005, 2010, 2012) 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 (2013) 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 1981; Banks and Sundaram 1992), bandits where an adversary has control over the payoffs (Auer et al. 2002/03), bandits with dependent arms (Pandey, Chakrabarti, and Agarwal 2007), bandits where multiple arms can be chosen at the same time (Whittle 1988), bandits whose arms yield rewards even when they are inactive (Glazebrook, Kirkbride, and Ruiz-Hernandez 2006), and bandits with switching costs (Banks and Sundaram 1994).

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 (1988). 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 2008; Veatch and Wein 1996; Whittle 1988; Faihe and Müller 1998; La Scala and Moran 2006). In aircraft surveillance, Ny, Dahleh, and Feron (2008) 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 (1996) 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

8. Basu, Bose, and Ghosh (1990), Bergemann and Välimäki (2006), and Mahajan and Teneketzis (2008) are excellent surveys of the literature on bandit models. The monographs by Presman and Sonin (1990), Berry and Fristedt (1985) and Gittins, Glazebrook, and Weber (2011) contain more detailed presentations.

9. Some bandits with switching costs can be modeled as restless bandits (Jun 2004).

an inventory increases holding costs. In robotics, Faihe and Müller (1998) 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 Optimality of stopping rules

For classical bandit models with one known and one unknown arm, the optimality of stopping rules is a well known result (Berry and Fristedt 1985; El Karoui and Karatzas 1994). Several approaches to establish this 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 (1987), allows a very simple formulation of the measurability constraints on the strategies. However, it is not well-adapted 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 1999; Keller, Rady, and Cripps 2005; Cohen and Solan 2013). However, in our model, 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, it is not clear a-priori if the value function is a solution in a classical sense. Pham (1995, 1998) showed that under suitable assumptions, the value function is a viscosity solution of the HJB equation. However, 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 2001).[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 unknown arm sooner rather than later. While this interchange argument was originally developed by Berry and Fristedt (1985) 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.

2.3 Indexability

In the classical bandit model, Gittins (1979) characterized optimal strategies 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 known and one unknown arm.

In general, optimal strategies in restless bandit models do not admit an index representation. However, a Lagrangian relaxation of the problem proposed by Whittle (1988) yields index strate-

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

gies that are approximately optimal (Weber and Weiss 1990, 1991). The corresponding "Whittle index" (Whittle 1988) 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 known arm is optimal is increasing in the reward from the known arm.[11]

The question of indexability of restless bandit models is subtle and not yet fully understood. Gittins, Glazebrook, and Weber (2011) 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 1975). In another line of research, Nino-Mora (2001) 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 bandit models – this is the subject of this paper.[12]

3 A Two-Armed Restless Bandit

3.1 Basic Building Blocks

Time $t \in [0, \infty)$ is continuous and there is one agent. Nature moves first and assigns a type $\Theta \in \{0, 1\}$ and an initial human capital H_0 to the agent.[13] The agent does not know her type but believes in being of the high type ($\Theta = 1$) with probability P_0. At each instant of time she decides what fraction of time to invest in the known and the unknown arm. Let $U_t \in [0, 1]$ be her investment decision at time t; $U_t = 1$ standing for investment in the unknown arm and $U_t = 0$ for investment in the known arm. Investments in the unknown arm increase her human capital, whereas investments in the known arm allow it to depreciate. The resulting level of human capital at time t is denoted by H_t. The rate at which human capital increases is denoted by $\alpha(1, H_t)$ and the rate at which it decreases $\alpha(0, H_t)$. Thus the human capital process solves the (pathwise) differential equation

$$dH_t = \big(U_t\alpha(1, H_t) + (1 - U_t)\alpha(0, H_t)\big)dt. \tag{1}$$

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.

12. Some results in this direction have been obtained for various bandit models related to sensor management, see the survey of Washburn (2008). Other classes of indexable problems are the dual speed problem of Glazebrook, Nino-Mora, and Ansell (2002), the maintenance models of Glazebrook, Ruiz-Hernandez, and Kirkbride (2006), and the spinning plates and squad models of Glazebrook, Kirkbride, and Ruiz-Hernandez (2006). 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 assume perfect information.

13. To fix ideas, we refer to H as the human capital of the agent, but our model is not bound to this interpretation.

At each instant of time, the agent receives a random reward dR_t that is characterized by the following three quantities:

$$\begin{aligned} b_t &= U_t\beta(\Theta, H_t) + (1 - U_t)k, && \text{(drift)} \\ c_t &= U_t\sigma(H_t)^2, && \text{(volatility)} \\ \nu_t(dr) &= U_t K(\Theta, H_t, dr) && \text{(jump measure)} \end{aligned} \tag{2}$$

Vaguely speaking, equations (2) describe the drift, volatility and jump measure of the reward process;

$$B_t = \int_0^t b_s ds, \quad C_t = \int_0^t c_s ds, \quad \mu(dt, dx) = \nu_t(dx)dt \tag{3}$$

are its semimartingale characteristics. These characteristics are defined with respect to some truncation function $\chi : \mathbb{R} \to \mathbb{R}$ that we fix once and for all. χ is equal to the identity on a neighborhood of zero and bounded, continuous. If the jump measure $K(\theta, h, dr)$ is finite, the truncation function is not needed and can be set to zero for all purposes.

If follows that the payoff to the known arm is deterministic and equals $k\,dt$. The payoff to the unknown arm is random and depends on the type and the level of human capital.[14] Rewards are discounted with a discount rate $\rho > 0$, and the agent tries to maximize her expected future rewards

$$\mathbb{E}\left(\int_0^\infty \rho e^{-\rho t} dR_t\right). \tag{4}$$

Only the reward of the arm that is chosen is observable. Formally, the coefficients $\alpha, \beta, \sigma, K, k, \rho$ of the model and the history of the processes U, H, R are known, but Θ is not.[15] Investment decisions U_t are restricted to depend on available information. Thus the agent's problem is a control problem with partial observations. A precise formulation of the problem is developed in section 3.3. When $\alpha = 0$ or b, c, ν do not depend on human capital, the model reduces to a classical bandit model. Otherwise, it is a bandit with evolving arms or restless bandit.

To assure that the control problem is well-defined, we make the following assumptions on the coefficients of the model.

Assumption 1 (Boundedness and regularity)**.** *The functions*

$$\alpha, \beta : \{0, 1\} \times \mathbb{R} \to \mathbb{R}, \quad \sigma : \mathbb{R} \to \mathbb{R} \tag{5}$$

are measurable and K is a transition kernel from $\{0, 1\} \times \mathbb{R}$ to $\mathbb{R} \setminus \{0\}$. Furthermore, the expressions

$$\alpha(u, h), \beta(\theta, h), \sigma(h), \int_{\mathbb{R}} \left(|r|^2 \wedge |r|\right) K(\theta, h, dr) \tag{6}$$

are uniformly bounded and $\alpha(u, h)$ is Lipschitz continuous in h.

14. The volatility does not depend on the type because the inference problem would be trivial in that case.
15. This is a key point of departure from Glazebrook, Kirkbride, and Ruiz-Hernandez (2006).

The uniform bounds on the coefficients and the Lipschitz condition ensure that human capital process is well-defined and that the expectation in equation (4) exists. The integrability condition on the jump measure implies that the reward process is a special semimartingale. This means that it has a compensator, which is a predictable process of integrable variation that differs from the reward process by a local martingale. Intuitively, the existence of a compensator means that agents are able to form expectations about the infinitesimal increments of the reward process.

3.2 Relation to classical bandit models

In this section, we relate our model to a number of important classical bandit models. In these models, the reward process is either Brownian motion with unknown drift (e.g. Bolton and Harris (1999)), a Poisson process with unknown intensity (e.g. Keller and Rady (2010)), or more generally, a Levy process with unknown Levy triple (e.g. Cohen and Solan (2013)). For consistency, we impose our notation on the models that we discuss in this section.

We begin with the simplest case where the reward process is Brownian motion with unknown drift. This model was introduced by Chernoff and Ray (1965) and subsequently extended by Bolton and Harris (1999, 2000). Using our notation, the reward process in this model is

$$dR_t = U_t\beta(\Theta)dt + (1-U_t)kdt + \sqrt{U_t}\sigma dW_t, \tag{7}$$

where W is Brownian motion independent of Θ. The square root in equation (7) permits an interpretation of $[0,1]$-valued controls U_t as fractions of time devoted to the unknown arm. Namely, the process $\int_0^t \sqrt{U_s}dW_s$ is equal in distribution to the time changed process W_{T_t}, where $T_t = \int_0^t U_s ds$ is the accumulated amount of investment in the unknown arm (Kallsen and Shiryaev 2002, theorem 1.5). The characteristics of the reward process in equation (7) are (B, C, μ) as in (3) with

$$b_t = U_t\beta(\Theta) + (1-U_t)k, \quad c_t = U_t\sigma^2, \quad \nu_t = 0. \tag{8}$$

The second case that we discuss is when the reward is a Poisson process with unknown intensity. The first treatment of this model, by Presman and Sonin (1990), is extended by Keller and Rady (2010). In their model, the probability that the agent receives a lump-sum payoff in the interval $[t, t+dt)$ is $U_t\lambda(\Theta)dt$, where λ is a non-negative function of Θ, and the distribution of lump-sum payoffs is given by a probability measure K on $\mathbb{R}\backslash\{0\}$. This statement can be made precise by interpreting it as a specification of the characteristics of the reward process. They are (B, C, μ) as in (3) with

$$b_t = (1-U_t)k, \quad c_t = 0, \quad \nu_t(dr) = U_t\lambda(\Theta)K(dr). \tag{9}$$

The last, and most general, case that we discuss are Lévy bandits (Cohen and Solan 2013). The reward process in their model is driven by a Lévy process X whose Lévy triplet depends on the unknown type Θ and is given by $(\beta(\Theta), \sigma, K(\Theta, \cdot))$. Under a strategy U, the agent obtains the

reward X_{T_t} from the unknown arm, where T is the time change

$$T_t = \int_0^t U_s ds. \tag{10}$$

The reward from the known arm is $k(t - T_t)$. The approach of defining reward processes using time changes is due to Mandelbaum (1987). It allows for a clean definition of admissible strategies and circumvents difficulties that arise when trying to add a strategy-dependent jump term to equation (7). An alternative to using time changes is to specify the characteristics of the reward process. They are given by (B, C, μ) as in equation (3) with

$$b_t = U_t\beta(\Theta) + (1 - U_t)k, \quad c_t = U_t\sigma^2, \quad \nu_t(dr) = U_t K(\Theta, dr). \tag{11}$$

Taken together, the classical bandit models presented in this section can be formulated in a convenient setting using semimartingale characteristics. This demonstrates that the only difference between these models and ours is that the characteristics in our model depend not only on the type of the agent, but also on an additional variable quantifying the amount of past investment.

3.3 The partially observable (p.o.) control problem

In this section, we describe: (a) what it means for the reward process R to be controlled by U and (b) what it means for U to be non-anticipative and to depend only on observable quantities. In an effort to make the paper self-contained, we err on the side of providing more detail.

The first issue (a) is straightforward. As discussed in the previous sections, we found it convenient to specify the distribution of the reward process by its semimartingale characteristics. It is well-known that specifying characteristics (B, C, μ) for R is equivalent to the following martingale problem (see Jacod and Shiryaev (2003, theorem II.2.42)): for each $f \in C_b^2(\mathbb{R})$, the process

$$f(R) - f(R_0) - f'(R_-) \cdot B - \frac{1}{2} f''(R_-) \cdot C - \big(f(R_- + r) - f(R_-) - f'(R_-)\chi(r)\big) * \mu \tag{12}$$

is a martingale. (We use the notation $\cdot$ and $*$ of Jacod and Shiryaev (2003) to denote stochastic integration with respect to semimartingales and random measures.) This brings us closer to the stance of controlled Markov processes. The reward process itself is not Markov, but the three-dimensional state-observation process (Θ, H, R) is, at least under constant control. It is well known that Markov processes can be specified by their generator. In our case, this would be a non-local second order differential operator. The coefficients of this operator are closely related to the drift, volatility and jump measure defined in equation (2). This puts our model in the framework of partially observed controlled Markov processes.[16] Yet another option is to formulate an SDE for the reward process. To account for jumps, one could try to add a term dN_t to the right-hand side of equation (7), where N is a compound Poisson process whose jump intensity depends on the type

16. See also Kurtz and Ocone (1985) and Kurtz and Stockbridge (1998)) who introduced a filtered martingale problem characterizing the conditional distribution of the unobserved state process.

and the control. It follows that one needs to consider a family of compound Poisson processes to allow for arbitrary controls. Then the equation turns into an SDE driven by non-linear Lévy noise in the sense of Kolokoltsov (2010, theorems 3.7 and 3.11).

We now come to the second task (b) of defining what it means for U to be non-anticipative and to depend only on observable quantities. This matter is rarely discussed in the applied bandit literature. However, the theoretical literature shows that it can a delicate issue. A basic requirement is that the control process must be predictable with respect to the filtration generated by the observation process. Ceci and Gerardi (1998) demonstrated that this requirement is sufficient when the observation process is a counting process. Their result applies for example to the Poisson and exponential bandit models of Keller, Rady, and Cripps (2005) and Keller and Rady (2010).

However, the above requirement is not stringent enough in general. For example, in our model, it would be tempting to admit any control process U that is $\mathbb{F}^{H,R}$-predictable, where $\mathbb{F}^{H,R}$ is the filtration generated by the observable processes H and R. (Generally, we will write $\mathbb{F}^X = (\mathcal{F}_t^X)_{t\geq 0}$ for the filtration generated by a process X.) But the differential equation (1) for H shows that U can be reconstructed from H, at least when $\alpha \neq 0$ and under a continuity assumption like U being càglàd. Then $\mathbb{F}^U \subseteq \mathbb{F}^{H,R}$ holds automatically and it is pointless to require U to be $\mathbb{F}^{H,R}$-predictable. Namely, any càglàd process U is $\mathbb{F}^{H,R}$-predictable, regardless of whether it depends on the supposedly unobserved state or not. Similar problems arise when requiring U to be $\mathbb{F}^R$-predictable since U can be reconstructed from the quadratic variation process of R.[17]

We use an approach popular in the early works in optimal control theory (Fleming and Nisio (1966); Wonham (1968); Kohlmann (1982)). Namely, we require that $U_t(\omega) = U_t(\omega')$ holds whenever the reward process satisfies $R_s(\omega) = R_s(\omega')$ for all $s < t$. This is equivalent to defining the control process as a functional of the reward process in the sense that $U = F(R)$, where F is a predictable process on Skorokhod space $D(\mathbb{R})$. This space is the canonical space of càdlàg paths, which are right-continuous functions $[0, \infty) \to \mathbb{R}$ with left limits. It is natural to assume that every sample path of the reward process is an element of this space. Under the strong measurability condition on the control, we are not able to prove that the set of admissible controls is compact and have to establish the existence of optimal controls in an indirect way. To this aim, we will transform the partially observed problem into a problem of full observations called the separated problem. This will be done in section 3.4.

Before we can define control processes, we need to make precise in what way the law of the reward process is determined by its characteristics. We follow the notation of Jacod and Shiryaev (2003, chapter IV). $\mathbb{P}$ is called a solution of the martingale problem $\mathscr{s}(\mathcal{H}, R \mid \eta; B, C, \mu)$ if $\mathbb{P}$ is a probability measure on the filtered space $(\Omega, \mathcal{F}, \mathbb{F})$ on which R, B, C, μ are defined such that $\mathbb{P}$ coincides with η on $\mathcal{H} \subseteq \mathcal{F}_0$ and R is a càdlàg $(\mathbb{F}, \mathbb{P})$-semimartingale with characteristics (B, C, μ). Existence, uniqueness and local uniqueness of the martingale problem are defined as usual.

We now define control processes as predictable processes on Skorokhod space $D(\mathbb{R})$ and call

17. Moreover, putting the reward process into a control-independent form by Zakai's measure transform (Fleming and Pardoux 1982) is not possible for general bandit models.

them admissible if a certain martingale problem associated to them is well-posed.[18] Well-posedness of the martingale problem is exactly what is needed to establish the equivalence to the separated problem, as can be seen from the proof of lemma 2 (in Appendix A) where the filtering equation is derived.

Definition 1 (Admissible control process)**.** *Let R be the coordinate process on Skorokhod space $(D(\mathbb{R}), \mathcal{F}, \mathbb{F})$. A predictable $[0,1]$-valued process F on this space is called an admissible control process if for all $\theta \in \{0,1\}$ and $h_0, r_0 \in \mathbb{R}$, existence and local uniqueness holds for the martingale problem $\mathcal{A}(\mathcal{F}_0, R \mid \delta_{r_0}; B, C, \mu)$ on $D(\mathbb{R})$, where B, C, μ, H are defined via equations* (1), (2), (3) *with $\Theta = \theta$, $U = F$, and $H_0 = h_0$.*

The next definition states that a p.o. control is a probability space endowed with U, Θ, H, R as in the previous section such that the control can be written as $U = F(R)$ for an admissible control process F. The conditions $P_0 = p_0$ and $H_0 = h_0$ are called the initial conditions of the control.

Definition 2 (Partially observable control)**.** *Let $\mathcal{B} = (\Omega, \mathcal{F}, \mathbb{F}, \mathbb{P})$ be a stochastic basis endowed with a $\{0,1\}$-valued random variable Θ, a continuous process H and a càdlàg process R. Furthermore, let F be an admissible control process and let $U = F(R)$ be the composition of F with R. If $\mathbb{P}$ solves the martingale problem $\mathcal{A}(\mathcal{F}_0, R \mid \eta; B, C, \mu)$, where B, C, μ, H, η are defined via equations* (1), (2), (3) *and*

$$\eta(\Theta = 1) = 1 - \eta(\Theta = 0) = p_0, \quad \eta(H_0 = h_0) = \eta(R_0 = 0) = 1, \tag{13}$$

then we call the tuple $\mathcal{U} = (\mathcal{B}, U, \Theta, H, R, p_0, h_0)$ a partially observable (p.o.) control with initial conditions (p_0, h_0). We write $\mathbb{U}^{p.o.}_{p_0,h_0}$ for the set of all such $\mathcal{U}$.

The set of p.o. controls is not empty. Indeed, any deterministic control process is admissible. To see this, note that the characteristics (B, C, μ) are deterministic in this case. It is well-known that this implies existence and local uniqueness of the martingale problem, see Jacod and Shiryaev (2003, theorem III.2.16). We will now define the value of a p.o. control.

Definition 3 (Value of p.o. control)**.** *The value of a p.o. control $\mathcal{U} \in \mathbb{U}^{p.o.}_{p_0,h_0}$ is*

$$J^{p.o.}(\mathcal{U}) = \mathbb{E}\left(\int_0^\infty \rho e^{-\rho t} dR_t\right), \tag{14}$$

where the expectation is taken with respect to $\mathbb{P}$. The value function for the p.o. control problem is

$$V^{p.o.}(p_0, h_0) = \sup\left\{J^{p.o.}(\mathcal{U}) : \mathcal{U} \in \mathbb{U}^{p.o.}_{p_0,h_0}\right\}. \tag{15}$$

18. Klein and Rady (2011) requires a similar condition in a multi-agent setting and gives it the interpretation that the solution can be obtained as a discrete-time limit. This interpretation is consistent with our definition since piecewise constant control processes are admissible. Friedman and Yavin (1980) and Kohlmann (1982) make continuity assumptions on controls to guarantee the well-posedness of the associated martingale problem. However, these stronger assumptions are not necessary.

Note that it follows from the bounds in assumption 1 that any control has finite and well-defined value, see also lemma 1 below.

We end this section with a remark on the interpretation of $[0,1]$-valued as opposed to $\{0,1\}$-valued controls. The interpretation of U_t as the fraction of time devoted to the unknown arm has already been given in the discussion of time changes in section 3.2. This interpretation is possible because of the linearity of the drift, volatility and jump measure in the control (see equations (2)). Another interpretation of U_t is as a relaxed, i.e., measure valued or randomized, control. The measure associated to U_t is $U_t\delta_1(du) + (1-U_t)\delta_0(du)$, where δ_1 and δ_0 are Dirac measures on $\{0,1\}$. In general, one has to work with relaxed controls to get good existence results for optimal strategies. It turns out that in our model, the additional generality provided by relaxed controls is not necessary. Indeed, we will prove the existence of optimal non-relaxed controls in theorem 2.

3.4 The separated (se.) control problem

The p.o. control problem is modified in several ways to solve it. The first step is to transform it into a so-called separated (se.) problem, which is standard in control theory. To derive the se. from the p.o. problem, the unobserved type Θ is replaced by its filter P_t. The filter is the conditional distribution of Θ given the observations up to time t. Since Θ is $\{0,1\}$-valued, P_t can be represented as a real number $P_t \in [0,1]$. In economic terms, P_t is the posterior belief of the agent in being of the high type. In the se. problem, the agent controls the fully observed process (P,H,R). Under constant controls, this process is Markov, see Kurtz (1998). Under some regularity assumptions on the generator of (P,H,R), it can be shown that optimal Markov strategies for the se. problem exist (see Kurtz and Stockbridge (1998, 1999)).

The second step is to reduce the dimension of the state space by one. This is made possible by the fact that the evolution of the reward process R depends on the values of Θ and H, but not on the value of R itself. This can be seen from the characteristics of the reward process in equation (2). Our model inherits this structure from classical bandit models where the evolution of the reward process also does not depend on its current value. We will now explain how the state variable R can be eliminated. The details of the argument can be found in the proof of lemma 2. First, note that $\mathbb{F}^R = \mathbb{F}^{H,R}$ because H is $\mathbb{F}^U$-adapted and U is $\mathbb{F}^R$-adapted. Taking $\mathbb{F}^R$-optional projections, one obtains that the $\mathbb{F}^R$-characteristics of R can be expressed via (P,H), which is the $\mathbb{F}^R$-optional projection of (Θ,H). Then also the characteristics of (P,H), which were originally expressed in terms of the characteristics of (P,H,R), can be expressed in terms of (P,H). It remains to express the value of a strategy using (P,H) instead of (P,H,R). This can be done by replacing R by its $\mathbb{F}^R$-compensator.

It remains to derive an equation for the filter P. For a p.o. control with $0 < p_0 < 1$ and $\theta \in \{0,1\}$, let $\mathbb{P}_\theta$ be the measure $\mathbb{P}$ conditioned on $\Theta = \theta$. It seems natural to assume that a finite amount of experimentation with the unknown arm cannot fully reveal the type Θ. If this is the case, then $\mathbb{P}_0$ and $\mathbb{P}_1$ are equivalent on $\mathcal{F}_t^R$ for all t. We will now explore some consequences of this assumption. By Girsanov's theorem, see Jacod and Shiryaev (2003, theorem III.3.24), there exist

measurable functions $\phi_1 : \mathbb{R} \to \mathbb{R}, \phi_2 : \mathbb{R}^2 \to (0, \infty)$ such that the following relations hold and are well-defined.

$$\begin{aligned} \beta(1,h) = \beta(0,h) + \sigma(h)^2\phi_1(h) + \int_{\mathbb{R}} \big(\phi_2(h,r) - 1\big)\chi(r)K(0,h,dr), \\ K(1,h,dr) = \phi_2(h,r)K(0,h,dr) \end{aligned} \tag{16}$$

The characteristics of the reward process for high and low type agents agree if and only if equation (16) holds with $\phi_1 = 0$ and $\phi_2 = 1$. Vaguely speaking, ϕ_1 accounts for differences in the drift and ϕ_2 for differences in the jumps. This is most clearly seen in the compound Poisson case where the jump measures $K(\theta, h, dr)$ are finite and the truncation function χ can be set to zero. Then

$$\phi_1(h) = \frac{\beta(1,h) - \beta(0,h)}{\sigma(h)^2}, \quad \phi_2(h,r) = \frac{K(1,h,dr)}{K(0,h,dr)} \tag{17}$$

holds. Thus ϕ_1 is the volatility-adjusted difference between the drifts and ϕ_2 is the Radon-Nikodym derivative of the high type over the low type jump measure. This means that jumps of hight r are $\phi_2(h,r)$ more likely for high type agents than for low type agents, holding the level of human capital fixed at h. We will see in section 3.7 that ϕ_1 and ϕ_2 also quantify the informativeness of the rewards from the unknown arm.

To be able to conclude a-priori that the measures $\mathbb{P}_1$ and $\mathbb{P}_0$ are locally equivalent with respect to the filtration $\mathbb{F}^R$, we make a boundedness assumption on ϕ_1, ϕ_2. This will also allow us to derive the filtering equation (see the proof of lemma 2). The exact form of the bounds stems from Cheridito, Filipović, and Yor (2005, theorem 2.4) or Lépingle and Mémin (1978, Théorème IV.3).

Assumption 2 (Local equivalence of measures)**.** *There exist functions ϕ_1, ϕ_2 satisfying* (16) *such that the expressions*

$$\phi_1(h), \quad \int_{\mathbb{R}} \big(\phi_2(h,r)\log\phi_2(h,r) - \phi_2(h,r) + 1\big)K(0,h,dr) \tag{18}$$

are locally bounded in h.

We now give the definition of the se. control problem.

Definition 4 (Separated control)**.** *Let $\mathcal{B} = (\Omega, \mathcal{F}, \mathbb{F}, \mathbb{P})$ be a stochastic basis endowed with a predictable $[0,1]$-valued process U, a càdlàg $[0,1]$-valued process P and a continuous $\mathbb{R}$-valued process H such that H satisfies equation* (1) *and such that P is an $(\mathbb{F}, \mathbb{P})$-semimartingale with characteristics (B, C, μ) as in equation* (3) *with*

$$b = -U\int_{\mathbb{R}} \big(q - \chi(q)\big)(j_*\overline{K})(P_-, H, dq), \quad c = UP_-^2(1-P_-)^2\phi_1(H)^2\sigma(H)^2, \quad \nu = U(j_*\overline{K})(P_-, H, dq). \tag{19}$$

Here, the measure $j_*\overline{K}$ is defined by

$$\forall g \in C_b(\mathbb{R}) : \int_{\mathbb{R}} g(q)(j_*\overline{K})(p,h,dq) = \int_{\mathbb{R}} g\big(j(p,h,q)\big)\overline{K}(p,h,dq), \tag{20}$$

$$\overline{K}(p,h,dq) = pK(1,h,dq) + (1-p)K(0,h,dq), \quad j(p,h,q) = \frac{p\phi_2(h,q)}{p\phi_2(h,q)+1-p} - p. \tag{21}$$

If furthermore the initial conditions

$$\mathbb{P}(P_0 = p_0) = \mathbb{P}(H_0 = h_0) = \mathbb{P}(R_0 = 0) = 1 \tag{22}$$

hold, then we call the tuple $\mathcal{U} = (\mathcal{B}, U, P, H, p_0, h_0)$ a separated control with initial conditions (p_0, h_0) and we write $\mathbb{U}^{se.}_{p_0,h_0}$ for the set of all such $\mathcal{U}$.

Remark (Properties of P)**.** Note that (19) implies that P is a local martingale, see Jacod and Shiryaev (2003, proposition II.2.29). It is also bounded, so it is a martingale. Notice how the volatility and jumps of P go to zero as P approaches the boundary of $[0,1]$. Therefore the condition that P is $[0,1]$-valued is satisfied automatically, at least under some additional regularity assumptions on the coefficients of the generator, see e.g. Simon (2000); Buckdahn et al. (2010); Filipovic, Tappe, and Teichmann (2012).

Definition 5. *The value of a se. control $\mathcal{U} \in \mathbb{U}^{se.}_{p_0,h_0}$ is*

$$J^{se.}(\mathcal{U}) = \mathbb{E}\left(\int_0^\infty \rho e^{-\rho t} \left(\overline{\beta}(P_{t-}, H_t)U_t + k(1-U_t) + U_t \int \big(r - \chi(r)\big)\overline{K}(P_{t-}, H_t, dr) \right) dt \right), \tag{23}$$

where

$$\overline{\beta}(p,h) = p\beta(1,h) + (1-p)\beta(0,h) \tag{24}$$

and where the expectation is taken with respect to the measure $\mathbb{P}$. The value function for the separated control problem is

$$V^{se.}(p_0,h_0) = \sup\left\{ J^{se.}(\mathcal{U}) : U \in \mathbb{U}^{se.}_{p_0,h_0} \right\}. \tag{25}$$

3.5 Equivalence of the p.o. and se. problem

In this section, we will prove the equivalence between the p.o. and the se. problem. This will be our first result. It will be established by discretization in time (but not in space). To prove that se. controls can be approximated by step controls, we need the following assumptions.

Assumption 3 (Well-posedness of the filtering equation under deterministic control)**.** *For any deterministic process U and corresponding human capital process H satisfying equation* (1)*, existence and local uniqueness holds for the martingale problem $\mathscr{s}(\mathcal{F}_0, P \mid \eta; B, C, \mu)$ under any initial condition η. Here, P is the coordinate process on $D(\mathbb{R})$ and (B, C, μ) satisfy equation* (3) *with (b, c, ν) as in equation* (19)*. Furthermore, the process P is a.s. $[0,1]$-valued under the solution measure.*

Assumption 4 (Continuity of the coefficients)**.** *The expressions*

$$\beta(\theta, h), \quad \sigma(h), \quad \int_{\mathbb{R}} g(r) K(\theta, h, dr) \tag{26}$$

are continuous in h for all $\theta \in \{0, 1\}$ and all functions $g \in C_b(\mathbb{R})$ vanishing near the origin.

Theorem 1 (Separation theorem)**.** *Let assumptions 1–4 hold. Then the value functions of the separated and the partially observed problem are finite and agree:*

$$V(p_0, h_0) := V^{p.o.}(p_0, h_0) = V^{se.}(p_0, h_0) < \infty. \tag{27}$$

To establish the theorem, we use step controls that we define next. These are controls that are constant on subsequent time intervals of a fixed length $\delta > 0$.

Definition 6 (Step controls)**.** *A step process with step size $\delta > 0$ is a (càdlàg or càglàd) process that is constant on all intervals (t_i, t_{i+1}), where $t_i = \delta i$ for $i \in \mathbb{N}$. $\mathbb{U}^{p.o.,\delta}_{p_0,h_0}$ is the set of p.o. controls with initial condition (p_0, h_0) whose control process U is a $\{0, 1\}$-valued step process with step size δ. The value function corresponding to this class of controls is denoted by $V^{p.o.,\delta}(p_0, h_0)$. $\mathbb{U}^{se.,\delta}_{p_0,h_0}$ and $V^{se.,\delta}(p_0, h_0)$ are defined correspondingly.*

The separation theorem follows from a sequence of lemmas that can be found in the appendix. We now give a verbal proof of the theorem, highlighting the role that each individual lemma plays.
Proof. *Step 1.* All value functions are well-defined and finite by lemma 1.

Step 2. For every p.o. control, one can define a process $P = \mathbb{E}(\Theta|\mathcal{F}^R_t)$ as the conditional expectation of the type Θ given the past observations. It is shown in lemma 2 that P satisfies the filtering equation (19) in the definition of se. controls. It follows that every p.o. control can be interpreted as a se. control. Moreover, the p.o. and se. control have the same value. By taking the supremum over all controls or step controls, one obtains that

$$V^{\text{p.o.}}(p_0, h_0) \leq V^{\text{se.}}(p_0, h_0), \quad V^{\text{p.o.},\delta}(p_0, h_0) \leq V^{\text{se.},\delta}(p_0, h_0). \tag{28}$$

Step 3. It remains to construct for every se. control a p.o. control of at least the same value. By a standard argument in lemma 4, se. controls can be approximated arbitrarily well by se. step controls. Formally, this is expressed by the equation

$$\sup_{\delta} V^{\text{se.},\delta}(p_0, h_0) = V^{\text{se.}}(p_0, h_0). \tag{29}$$

Thus it is sufficient to show that every se. step control corresponds to a p.o. control of at least the same value. This is done in lemma 3. In this lemma, the p.o. control is constructed recursively for each step of the control process by stitching together solution measures to the p.o. martingale problem under constant control. This establishes the relation

$$V^{\text{se.},\delta}(p_0, h_0) \leq V^{\text{p.o.},\delta}(p_0, h_0) \tag{30}$$

Together with the results of the previous step this immediately implies that equality holds in (30). By allowing arbitrarily small step sizes δ one obtains that the se. and p.o. value functions are finite and agree:

$$V^{\text{p.o.}}(p_0, h_0) \leq V^{\text{se.}}(p_0, h_0) = \sup_{\delta} V^{\text{se.},\delta}(p_0, h_0) \leq \sup_{\delta} V^{\text{p.o.},\delta}(p_0, h_0) \leq V^{\text{p.o.}}(p_0, h_0) \tag{31}$$

■

3.6 Equivalence to optimal stopping

The next step is to reduce the stochastic control problem to an equivalent stopping problem. That is, there are optimal stopping controls as defined below.

Definition 7. *A control $\mathcal{U}$ is called a stopping control if the associated control process U is of the form $U_t = \mathbb{1}_{[\![0,T]\!]}(t)$ for a stopping time T.*

The result hinges on the monotonicity condition that the expected rewards of the unknown arm increase while the arm is operated and decrease otherwise. Since the amount of past investment in the unknown arm is represented by the level of human capital, this is equivalent to assuming that the average reward from investments in the unknown arm is an non-decreasing function of the level of human capital. The infinitesimal version of this assumption can be stated as follows:

Assumption 5 (Monotonicity condition)**.** *The relation*

$$\alpha(0, h) \leq 0 \leq \alpha(1, h) \tag{32}$$

holds for all $h \in \mathbb{R}$ and the expression

$$\beta(\theta, h) + \int_{\mathbb{R}} \big(r - \chi(r)\big) K(\theta, h, dr) \tag{33}$$

is non-decreasing in $\theta, h \in \{0, 1\} \times \mathbb{R}$.

We are now ready to state our main theorem. Recall that $V = V^{\text{p.o.}} = V^{\text{se.}}$, see theorem 1.

Theorem 2 (Optimal stopping)**.** *Let assumptions 1–5 hold. Then the following stopping times T^* are optimal.*

1. $T^* = \inf\{t \geq 0 : V(P_t, H_t) \leq k\}$.
2. $T^* = \inf\{t \geq 0 : G(P_t, H_t) \leq k\}$, *where G is defined by*

$$G(p_0, h_0) := \inf\left\{ s : \sup_T \mathbb{E}\left(\int_0^T \rho e^{-\rho t}(dR_t - s dt) \right) \leq 0 \right\} = \sup_T \frac{\mathbb{E}\left(\int_0^T \rho e^{-\rho t} dR_t \right)}{\mathbb{E}\left(\int_0^T \rho e^{-\rho t} dt \right)}. \tag{34}$$

The suprema in the formula above are over all $\mathbb{F}^{P,H}$-stopping times T and the processes P, H, and R are governed by the constant strategy $U_t = 1$ for all t.

G formally coincides with Gittins' index. However, the payoff distribution of inactive arms can evolve. This is not allowed in classical bandit models, which are the object of Gittins' (1979) theory. The first formula in equation (34) is a continuous time version of Weber's (1992) modification of Whittle (1980). The second formula is the continuous time version of the original formulation of the index by Gittins and Jones (1974). Some references for the continuous time setting are El Karoui and Karatzas (1994, 1997); Bank and Küchler (2007).

An immediate consequence of theorem 2 is a characterization of optimal strategies by a curve that is typically referred to as the *decision frontier*.

Corollary 1. *There is a curve in the (p, h)-domain such that it is optimal to invest in the unknown arm if (P_t, H_t) lies to the right and above of the curve. Otherwise, it is optimal to invest in the known arm.*

Proof. The value function V is non-decreasing in its arguments by lemma 5 in the appendix 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. ■

Another consequence is the indexability of our bandit model in the sense of Whittle (1988). This means that the set of states in the (p, h)-domain where the known arm is optimal increases in the payoff k of the known arm. This property is obvious from equation (34).

Corollary 2. *Under assumptions 1–5, our restless bandit model is indexable in the sense of Whittle (1988). More generally, in a multi-armed bandit model, this holds for any arm that satisfies these assumptions.*

Theorem 2 follows from a sequence of lemmas. Its proof is short and can serve as a guide to how the lemmas are used. The core of the proof is lemma 7 where the optimality of stopping rules is shown. It is a generalization of an argument originally developed by Berry and Fristedt (1985, section 5.2) for classical bandits in discrete time. It turns out that our monotonicity assumption is exactly what is needed to make the proof work for restless bandits. Once it has been established that the problem is equivalent to optimal stopping, the characterization of optimal stopping times via the value function and via Gittins' index follow easily.

Proof of theorem 2. We make repeated use of the monotonicity assumption 5. A first consequence of this assumption is the monotonicity in (p_0, h_0) and convexity in p_0 of the value function. These properties are established in lemma 5. The result is then used in lemma 6 to prove a sufficient condition for the unique optimality of the unknown arm as an initial choice. Namely, when the expected immediate (myopic) payoff is higher for the unknown than for the known arm, then it is uniquely optimal to choose the unknown arm. This result is used in lemma 7 to prove that there exist optimal stopping rules for the discretized control problem. The argument is a modification of

Berry and Fristedt (1985, theorem 5.2.2) that allows the reward to depend on the level of human capital. Since it is already know from lemma 4 that controls can be approximated arbitrarily well by step controls, it follows that the set of p.o. and se. controls can be restricted to stopping controls without incurring any loss of value. Thus it has been established in the continuous time setting that the value function is the supremum over the values of stopping controls. However, it remains to show that optimal stopping controls exist. This is well-known for controlled Markov processes, so the result is established with ease in lemma 8 for the se. problem. By lemma 9, an optimal stopping rule for the se. problem yields also an optimal stopping rule for the p.o. problem. The characterization of optimal stopping controls as in theorem 2.1 is immediate from the proof of lemma 8. The equivalence to the formulas in theorem 2.2 is well-known, see e.g. Morimoto (1991, theorem 2.1) or El Karoui and Karatzas (1994, proposition 3.4). ■

3.7 Asymptotic learning

In this section, we calculate the limits of the belief process and formulate condition for its convergence to the true type. We cannot expect an agent to learn her true type if her accumulated investment over time is bounded. Formally, this is a consequence of assumption 2 implying the local equivalence of the probability measures $\mathbb{P}_1$ and $\mathbb{P}_0$ (see the proof of lemma 2). Here, for $\theta \in \{0,1\}$, $\mathbb{P}_\theta$ denotes the measure $\mathbb{P}$ conditioned on $\Theta = \theta$. On the other hand, an agent should be able to learn her true type in the limit if the amount she invests in the unknown arm goes to infinity. To show this, one needs to make an assumption on the minimum informativeness of the reward process about the type. The type can be learned from the drift and the jumps of the reward, but not from its volatility. Indeed, the volatility does not contain any information about the type, as can be seen from the definition of the characteristic triplet (b, c, ν) in equation (2). This is not a coincidence. Namely, by Girsanov's theorem, any dependence of the volatility on the type would reveal the type immediately, which would make the inference problem trivial.

The following assumption says that there is a lower bound on the difference between the drift and jump measure of the reward process for the high and the low type. Moreover, this bound has to be uniform in the level of human capital. Differences in the drift are quantified by the function ϕ_1 and differences in the jump probabilities by the function ϕ_2. These functions were defined and explained in section 3.4.

Assumption 6 (Lower bound on the informativeness of the reward process)**.**

$$\inf_h \frac{1}{8}\phi_1(h)^2\sigma(h)^2 + \frac{1}{2}\int_{\mathbb{R}} \left(1 - \sqrt{\phi_2(h,r)}\right)^2 K(0,h,dr) > 0 \tag{35}$$

Equation (35) is independent of the choice of the truncation function χ, which can be seen from Jacod and Shiryaev (2003, II.2.24). In the compound Poisson case where the truncation function

can effectively be set to zero, it takes the form

$$\inf_h \frac{1}{8}\left(\frac{\beta(1,h)-\beta(0,h)}{\sigma(h)}\right)^2 + \frac{1}{2}\int_{\mathbb{R}}\left(1-\sqrt{\frac{K(1,h,dr)}{K(0,h,dr)}}\right)^2 K(0,h,dr) > 0. \tag{36}$$

The terms in this formula have an intuitive meaning. Either the drift or the jump probabilities of high and low type agents must differ from each other. Moreover, higher volatility of the reward process requires larger differences in the drift.

The exact form of the bound stems from the Hellinger process $h(\frac{1}{2})$ of the measures $\mathbb{P}_1$ and $\mathbb{P}_0$ with respect to the filtration $\mathbb{F}^R$. The Hellinger process can be used to characterize the mutual singularity of these measures on $\mathcal{F}^R$, see Jacod and Shiryaev (2003, chapter IV). The mutual singularity in turn is equivalent to convergence of the belief process P_t to the true state Θ, which we call asymptotic learning. We are now able to formulate and prove a necessary and sufficient condition for asymptotic learning.

Theorem 3 (Asymptotic learning)**.** *Let the initial belief be non-doctrinaire in the sense that* $0 < p_0 < 1$*. Under assumptions 1 and 2, agents cannot learn their true type in finite time. If assumption 6 holds in addition, then the agent learns her true type in the long-term limit if and only if the accumulated amount of investment in the unknown arm goes to infinity.*

Proof. The equivalence of the probability measures $\mathbb{P}_1$ and $\mathbb{P}_0$ under assumptions 1 and 2 is proven in lemma 2. This implies that the process P_t never reaches zero or one in finite time and proves the first statement of the theorem. The second statement can be expressed formally as

$$\{P_\infty = \Theta\} = \left\{\int_0^\infty U_t dt = \infty\right\} \quad \mathbb{P}\text{-a.s.} \tag{37}$$

We show this equality for a p.o. control $\mathcal{U} \in \mathbb{U}^{\text{p.o.}}_{p_0,h_0}$ with control process U. For $\theta \in \{0,1\}$ let $\mathbb{P}_\theta$ be the conditional measure of $\mathbb{P}$ given $\Theta = \theta$. Then $\mathbb{P} = p_0\mathbb{P}_1 + (1-p_0)\mathbb{P}_0$. We know from the proof of lemma 2 that $\mathbb{P}_1$ and $\mathbb{P}_0$ are equivalent on $\mathcal{F}^R_t$ for all $t \in \mathbb{R}$. Let D be the $\mathbb{F}^R$-density process of $\mathbb{P}_1$ with respect to $\mathbb{P}_0$. We obtain from equation (56) that

$$\begin{aligned} d\langle D^c, D^c\rangle_t = D^2_{t-} d\langle L^c, L^c\rangle_t = D^2_{t-}\phi_1(H_t)^2 d\langle R^c, R^c\rangle_t = D^2_{t-}\phi_1(H_t)^2 U_t \sigma(H_t)^2 dt, \\ \Delta D_t = D_{t-}\Delta L_t = D_{t-}\big(\phi_2(H_t, \Delta R_t) - 1\big). \end{aligned} \tag{38}$$

Therefore the jump measure of D is $(\mathbb{F}^R, \mathbb{P})$-compensated by the measure

$$\mu(dt, dx) = U_t \Big(D_{t-}\big(\phi_2(H_t,\cdot) - 1\big)\Big)_* K(0, H_t, dx)dt. \tag{39}$$

It follows by Jacod and Shiryaev (2003, Corollary IV.1.37) that the Hellinger process $h(\frac{1}{2})$ of $\mathbb{P}_1$

and $\mathbb{P}_0$ satisfies

$$\begin{aligned} h\left(\frac{1}{2}\right) &= \frac{1}{8}\frac{1}{D_-^2}\cdot\langle D^c, D^c\rangle + \frac{1}{2}\left(1-\sqrt{1+\frac{x}{D_-}}\right)^2 * \mu \\ &= \int_0^{\cdot} U_t\left(\frac{1}{8}\phi_1(H_t)^2\sigma(H_t)^2 + \frac{1}{2}\int_{\mathbb{R}}\left(1-\sqrt{\phi_2(H_t,x)}\right)^2 K(0,H_t,dx)\right)dt \end{aligned} \tag{40}$$

It follows from assumption 6 that $h(\frac{1}{2})_\infty = \infty$ if and only if $\int_0^\infty U_t dt = \infty$. Since $\mathbb{P}_1$ and $\mathbb{P}_0$ are locally equivalent, $0 < D_t < \infty$ holds for all $t \in \mathbb{R}$. Then Schachermayer and Schachinger (1999, theorem 1.5) implies that

$$\{h(\tfrac{1}{2})_\infty = \infty\} = \{D_\infty = 0\} \quad \mathbb{P}_0 - \text{a.s.} \tag{41}$$

This is equivalent to

$$\{h(\tfrac{1}{2})_\infty = \infty\} = \{P_\infty = 0\} \quad \mathbb{P}_0 - \text{a.s.} \tag{42}$$

where P is the $\mathbb{F}^R$-conditional expectation of Θ that is related to D by equation (57). Reversing the rôles of $\mathbb{P}_0$ and $\mathbb{P}_1$, one obtains that

$$\{h(\tfrac{1}{2})_\infty = \infty\} = \{P_\infty = 1\} \quad \mathbb{P}_1 - \text{a.s.} \tag{43}$$

Equations (42) and (43) establish the mutual singularity of $\mathbb{P}_1$ and $\mathbb{P}_0$ on $\mathcal{F}^R$ restricted to the set $\{h(\frac{1}{2})_\infty = \infty\}$. This proves the theorem. ■

The above result stands in contrast with the exponential bandits model of Keller, Rady, and Cripps (2005). In that model, jumps of the reward process are possible only for high type agents.[19] Observing a jump fully reveals the type. Thus agents in this model can learn their type in finite time.

Note that theorem 3 holds under any strategy, not just under the optimal one. However, it can be combined with theorem 2 to the following statement about asymptotic learning under optimal strategies: for any agent, asymptotic learning fails with positive probability. However, it takes place with positive probability for high type agents starting out above the frontier.[20]

3.8 Trajectories in the belief–human capital space

Let us restrict our attention to a range $(h_{\min}, h_{\max})$ of human capital that is invariant under the evolution of human capital. This means that the process H_t never leaves the interval $(h_{\min}, h_{\max})$ when H_0 lies in this interval. To exclude trivial cases, we also require the initial belief to be non-doctrinaire in the sense that $0 < p_0 < 1$ holds. We set out to describe the possible trajectories of an agent in belief–human capital space $(0,1)\times(h_{\min}, h_{\max})$ and to explore the importance of dynamic human capital in that context.

19. Thus assumption 2 is not satisfied in this model.

20. This can be compared to the linear network structure in Acemoglu et al. (2011, example 1.1) where asymptotic learning is guaranteed. This is because agents receive information at each stage regardless of whether they invest or not.

The results from the previous section together with the characterization of optimal strategies in terms of the decision frontier imply that agents meet one of two fates: they either remain above the frontier, in which case they learn their true type in the limit. Or they fall below the frontier at some point, in which case they cannot learn their true type. A first consequence of this observation is that in the long run, only high type agents are susceptible to making suboptimal investment decisions. Namely, they might drop down below the frontier because of bad luck and stop learning about their type. In contrast, all low type agents eventually choose the option they would also choose if they knew their type. (Depending on the parameters of the model and on the level of human capital of the agent, this might mean investing or not investing.) It follows that in the long run, compared to a setting with full information, agents invest too little in the unknown arm. This points to the importance of policies designed to increase investment in the unknown arm.

The effect of dynamic as opposed to static human capital on optimal investment is best seen by looking at the limit (P_∞, H_∞) of the belief–human capital process as time goes to infinity. This limit always exists as an element of $[0,1] \times [h_{\min}, h_{\max}]$ because P is a bounded martingale and H is an integral curve to vector field (at least after a stopping time where the agent might change her investment strategy).

In the case where human capital is dynamic in the strict sense that $\alpha(0,h) < 0 < \alpha(1,h)$ holds for $h \in (h_{\min}, h_{\max})$, agents never converge to the frontier. Instead, their human capital converges either to $h_{\min}$ or $h_{\max}$, depending on whether they stop investing in the unknown arm at some point in time or not. However, in the static case where $\alpha(1,h) = \alpha(0,h) = 0$ holds and human capital is constant, agents above the frontier hit the frontier with positive probability and remain there forever.

This consideration suggests that agents in the static human capital case accumulate at the frontier. This statement can be given a precise meaning. Assume that there is a population of agents whose initial belief and human capital is uniformly distributed. 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 the agents in the belief–human capital domain evolves over time and converges to the distribution of (P_∞, H_∞). Figures 1 and 2 depict the decomposition of this distribution into absolutely continuous and singular parts with respect to the Lebesgue measure. Singular parts are highlighted. They correspond to areas of vanishing Lebesgue measure where nevertheless, there is a positive fraction of agents in the long term limit. Thus highlighted areas in the graphs can be interpreted as accumulation points of agents. Notice that the frontier is highlighted in exactly the cases where human capital is constant. In all other cases, it is not highlighted and in fact, never occurs as a limit.

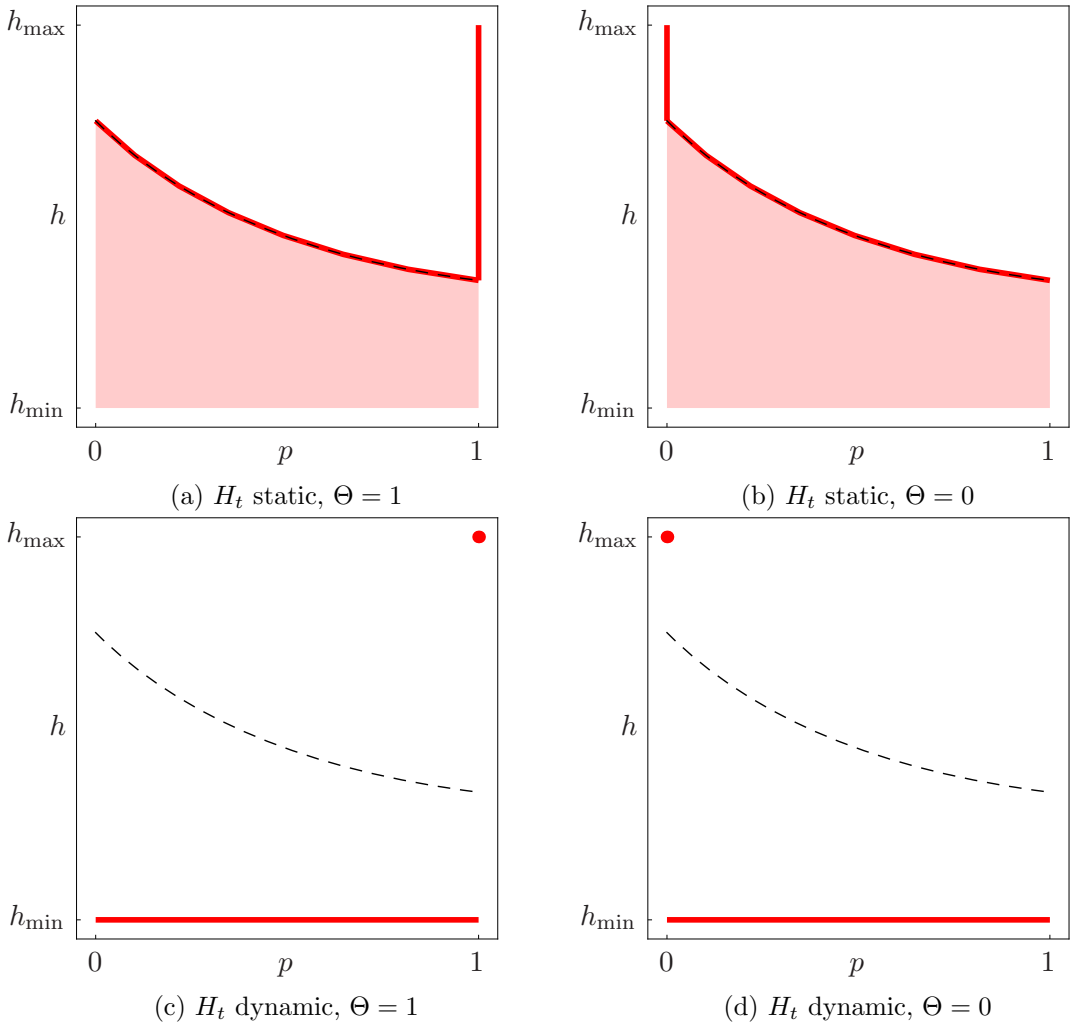
Figures 1 and 2 differ by the position of the frontier relative to the boundary of the belief–human capital domain. In figure 1, human capital can be so high (low) that the unknown (known) arm is optimal for all agents, regardless of their type. In other words, human capital has a stronger influence than the type. By contrast, in figure 2, the type has a stronger influence than human

capital, which makes the unknown (known) arm optimal for all high (low) type agents, regardless of their level of human capital. The dominance of human capital or the type is determined by the parameters of the model.

4 Conclusion

We discovered a class of restless bandit models of investment under uncertainty where payoffs are allowed to depend on the history of past investments in a monotonic way. We argue that this dynamic dependence is a defining feature of many economically important activities such as human capital formation or job search. Agents in our model have imperfect information and learn through observations of the reward process, which we allow to be a general semimartingale. We solve the model by showing that stopping rules are optimal and can be characterized by an index that formally coincides with Gittins' index. Moreover, we characterize the learning process by giving necessary and sufficient conditions for asymptotic learning.

Allowing arms in a bandit to evolve as in our model results in a stark empirical prediction – there are very few truly "marginal agents" (in the classic sense). Instead, once agents stop investing they drift to the boundary. In this case, optimal policies to foster human capital or unemployment policies designed to keep individuals in the workforce may have very different characteristics than optimal policy in standard life-cycle human capital or job search models. For instance, one could imagine that if policies can be targeted to individuals, optimal policy might wait until agents are close to the decision frontier and then subsidize investment in the risky arm if the probability that they are high type is large enough. Conversely, if one has to make lumpy transfers (e.g. invest in a community or a school) then optimal policy may be more complicated. With a new indexable class of restless bandit models, there are many potential avenues of future research.



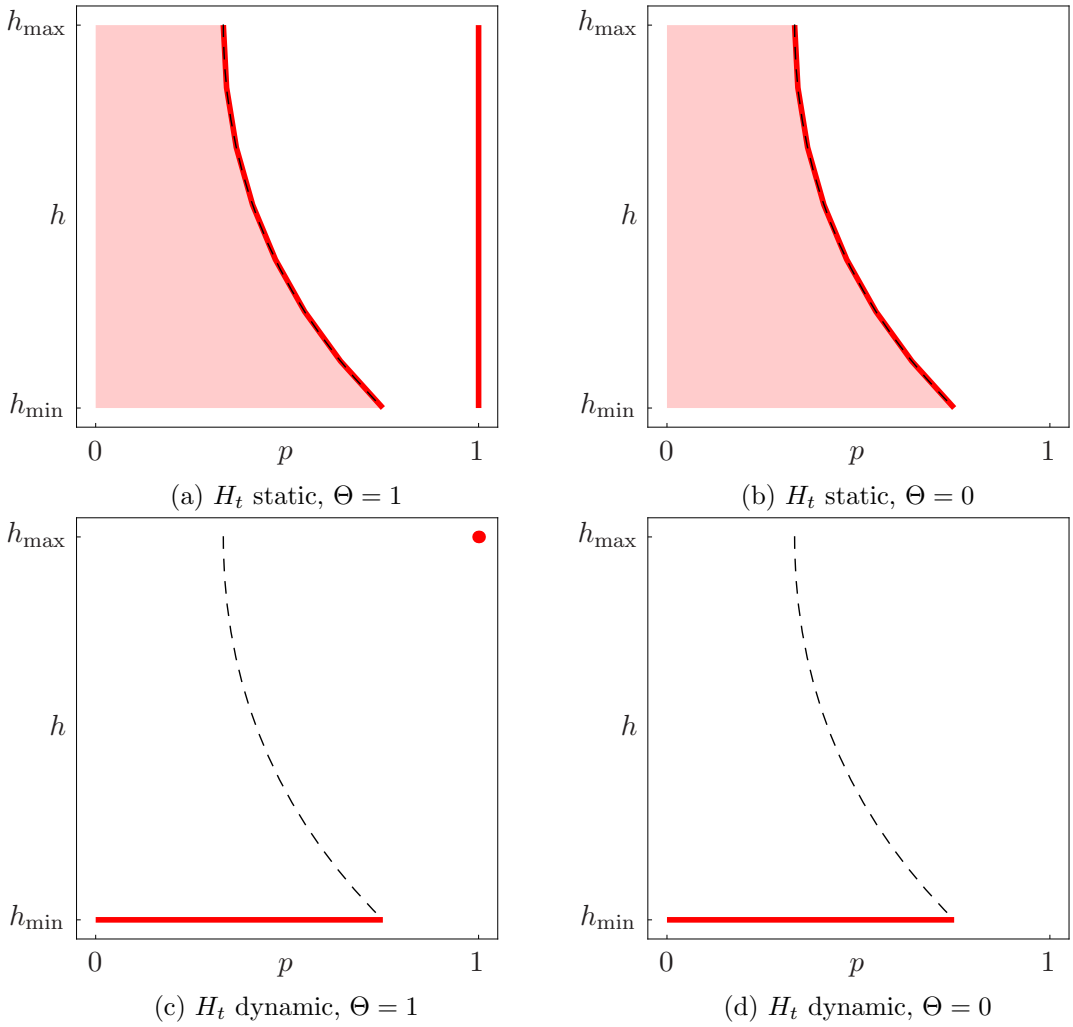
(a) H_t static, $\Theta = 1$

(b) H_t static, $\Theta = 0$

(c) H_t dynamic, $\Theta = 1$

(d) H_t dynamic, $\Theta = 0$

Figure 1: Absolutely continuous (opaque areas) and singular (bold lines and dots) parts of the distribution of (P_∞, H_∞) when (P_0, H_0) is uniformly distributed. The dashed line is the decision frontier. It intersects the left and right boundary of the (p, h)-domain because the parameters are such that the effect of human capital dominates the effect of Θ. Note that agents accumulate at the frontier when H_t is constant but move away from it when H_t is dynamic.



(a) H_t static, $\Theta = 1$

(b) H_t static, $\Theta = 0$

(c) H_t dynamic, $\Theta = 1$

(d) H_t dynamic, $\Theta = 0$

Figure 2: Absolutely continuous (opaque areas) and singular (bold lines and dots) parts of the distribution of (P_∞, H_∞) when (P_0, H_0) is uniformly distributed. The dashed line is the decision frontier. It intersects the left and right boundary of the (p, h)-domain because the parameters are such that the effect of Θ dominates the effect of human capital. Note that agents accumulate at the frontier when H_t is constant but move away from it when H_t is dynamic.

References

Acemoglu, Daron, Munther Dahleh, Ilan Lobel, and Asuman Ozdaglar. 2011. "Bayesian Learning in Social Networks." *The Review of Economic Studies* 78, no. 4 (October): 1201.

Auer, Peter, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. 2002/03. "The nonstochastic multiarmed bandit problem." *SIAM J. Comput.* 32 (1): 48–77 (electronic).

Bank, P., and C. Küchler. 2007. "On Gittins' index theorem in continuous time." *Stochastic processes and their applications* 117 (9): 1357–1371.

Banks, Jeffrey S., and Rangarajan K. Sundaram. 1992. "Denumerable-Armed Bandits." *Econometrica* 60, no. 5 (September): pages.

———. 1994. "Switching Costs and the Gittins Index." *Econometrica* 62, no. 3 (May): pages.

Basu, A., A. Bose, and JK Ghosh. 1990. *An Expository Review of Sequential Design and Allocation Rules.* Technical report. Department of Statistics, Purdue University.

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

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

Bolton, Patrick, and Christopher Harris. 1999. "Strategic experimentation." *Econometrica* 67 (2): 349–374.

———. 2000. "Strategic Experimentation: The Undiscounted Case." In *Incentives, Organization and Public Economics. Papers in Honour of Sir James Mirrlees,* edited by Peter J. Hammond and Gareth D. Myles, 53–68. Oxford / New York: Oxford University Press.

Buckdahn, R., M. Quincampoix, C. Rainer, and J. Teichmann. 2010. "Another proof for the equivalence between invariance of closed sets with respect to stochastic and deterministic systems." *Bulletin des sciences mathematiques* 134 (2): 207–214.

Ceci, Claudia, and Anna Gerardi. 1998. "Partially observed control of a Markov jump process with counting observations: equivalence with the separated problem." *Stochastic Process. Appl.* 78 (2): 245–260.

Cheridito, Patrick, Damir Filipović, and Marc Yor. 2005. "Equivalent and absolutely continuous measure changes for jump-diffusion processes." *Ann. Appl. Probab.* 15 (3): 1713–1732.

Chernoff, Herman, and S. N. Ray. 1965. "A Bayes sequential sampling inspection plan." *Ann. Math. Statist.* 36:1387–1407. ISSN: 0003-4851.

Cohen, A., and E. Solan. 2013. "Bandit problems with Levy payoff processes." *Mathematics of Operations Research* 38, no. 1 (February): 92–107.

Cunha, Flavio, and James J Heckman. 2006. "Investing in our Young People."

Cunha, Flavio, James J Heckman, Lance Lochner, and Dimitriy V Masterov. 2006. "Interpreting the evidence on life cycle skill formation." *Handbook of the Economics of Education* 1:697–812.

Dobbie, Will, and Roland Fryer. 2011. *Getting beneath the veil of effective schools: Evidence from New York City.* Technical report. National Bureau of Economic Research.

El Karoui, N., and I. Karatzas. 1994. "Dynamic allocation problems in continuous time." *Ann. Appl. Probab.* 4 (2): 255–286.

———. 1997. "Synchronization and optimality for multi-armed bandit problems in continuous time." *Computational and Applied Mathematics* 16:117–152.

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

Filipovic, D., S. Tappe, and J. Teichmann. 2012. "Invariant manifolds with boundary for jump-diffusions." *ArXiv e-prints* (February). arXiv: 1202.1076.

Fleming, Wendell H. 1989. "Generalized solutions and convex duality in optimal control." In *Partial differential equations and the calculus of variations, Vol. I,* 461–471. Vol. 1. Progr. Nonlinear Differential Equations Appl. Birkhäuser Boston.

Fleming, Wendell H., and Makiko Nisio. 1966. "On the existence of optimal stochastic controls." *J. Math. Mech.* 15:777–794.

Fleming, Wendell H., and Étienne Pardoux. 1982. "Optimal control for partially observed diffusions." *SIAM J. Control Optim.* 20 (2): 261–285.

Friedman, M., and Y. Yavin. 1980. "Optimal control of partially observable jump diffusion processes." *Internat. J. Systems Sci.* 11 (3): 323–335.

Gittins, J. C., and D. M. Jones. 1974. "A dynamic allocation index for the sequential design of experiments." In *Progress in statistics (European Meeting Statisticians, Budapest, 1972),* 241–266. Colloq. Math. Soc. János Bolyai, Vol. 9. Amsterdam: North-Holland.

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

Gittins, John, Kevin Glazebrook, and Richard Weber. 2011. *Multi-armed Bandit Allocation Indices.* Wiley-Blackwell.

Glazebrook, KD, C. Kirkbride, and D. Ruiz-Hernandez. 2006. "Spinning plates and squad systems: policies for bi-directional restless bandits." *Advances in applied probability* 38 (1): 95–115.

Glazebrook, KD, J. Nino-Mora, and PS Ansell. 2002. "Index policies for a class of discounted restless bandits." *Advances in Applied Probability* 34 (4): 754–774.

Glazebrook, KD, D. Ruiz-Hernandez, and C. Kirkbride. 2006. "Some indexable families of restless bandit problems." *Advances in Applied Probability* 38 (3): 643–672.

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

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

Kallsen, J., and A.N. Shiryaev. 2002. "Time Change Representation Of Stochastic Integrals." *Theory Probab. Appl.* 46 (3): 522.

Karatzas, I. 1984. "Gittins indices in the dynamic allocation problem for diffusion processes." *The Annals of Probability:*173–192.

Keller, Godfrey, and Sven Rady. 2010. "Strategic experimentation with poisson bandits" [in English]. *Theoretical Economics* 5, no. 2 (May): 275–311.

———. 2012. *Breakdowns.* `http://epub.ub.uni-muenchen.de/14316/`.

Keller, Godfrey, Sven Rady, and Martin Cripps. 2005. "Strategic experimentation with exponential bandits." *Econometrica* 73 (1): 39–68.

Klein, Nicolas, and Sven Rady. 2011. "Negatively correlated bandits." *The Review of Economic Studies* 78 (2): 693–732.

Klimov, GP. 1975. "Time-sharing service systems. I." *Theory of Probability & Its Applications* 19 (3): 532–551.

Kohlmann, M. 1982. "Existence of optimal controls for a partially observed semimartingale." *Stochastic Processes and their Applications* 13 (2): 215–226.

Kolokoltsov, Vassili N. 2010. *Nonlinear Markov processes and kinetic equations.* xviii+375. Vol. 182. Cambridge Tracts in Mathematics. Cambridge: Cambridge University Press.

Kurtz, Thomas G. 1998. "Martingale problems for conditional distributions of Markov processes." *Electron. J. Probab.* 3 (9): 1–29.

Kurtz, Thomas G., and Daniel Ocone. 1985. "A martingale problem for conditional distributions and uniqueness for the nonlinear filtering equations." In *Stochastic differential systems (Marseille-Luminy, 1984),* 224–234. Vol. 69. Lecture Notes in Control and Inform. Sci. Berlin: Springer.

Kurtz, Thomas G., and Richard H. Stockbridge. 1998. "Existence of Markov controls and characterization of optimal Markov controls." *SIAM J. Control Optim.* 36 (2): 609–653 (electronic).

Kurtz, Thomas G., and Richard H. Stockbridge. 1999. "Erratum: "Existence of Markov controls and characterization of optimal Markov controls"." *SIAM J. Control Optim.* 37 (4): 1310–1311 (electronic).

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

La Scala, BF, and B. Moran. 2006. "Optimal target tracking with restless bandits." *Digital Signal Processing* 16 (5): 479–487.

Lépingle, Dominique, and Jean Mémin. 1978. "Sur l'intégrabilité uniforme des martingales exponentielles." *Z. Wahrsch. Verw. Gebiete* 42 (3): 175–203.

Li, C., and M.J. Neely. 2011. "Network utility maximization over partially observable markovian channels." In *Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks (WiOpt), 2011 International Symposium on,* 17–24. IEEE.

Mahajan, A., and D. Teneketzis. 2008. "Multi-armed bandit problems." *Foundations and Applications of Sensor Management:*121–151.

Mandelbaum, Avi. 1987. "Continuous multi-armed bandits and multiparameter processes." *Ann. Probab.* 15 (4): 1527–1556.

Mazliak, Laurent. 1993. "Mixed control problem under partial observation." *Appl. Math. Optim.* 27 (1): 57–84.

McCall, B. P., and J. J. McCall. 1987. "A Sequential Study of Migration and Job Search." *Journal of Labor Economics* 5 (4): 452–476.

Morimoto, Hiroaki. 1991. "On average cost stopping time problems." *Probab. Theory Related Fields* 90 (4): 469–490.

Nino-Mora, J. 2001. "Restless bandits, partial conservation laws and indexability." *Advances in Applied Probability* 33 (1): 76–98.

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

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

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

Pham, Huyên. 1995. "Optimal stopping of controlled jump diffusion processes and viscosity solutions." *C. R. Acad. Sci. Paris Sér. I Math.* 320 (9): 1113–1118.

———. 1998. "Optimal stopping of controlled jump diffusion processes: a viscosity solution approach." *J. Math. Systems Estim. Control* 8 (1): 27 pp. (electronic).

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

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

Presman, Ernst L. 1990. "Poisson version of the two-armed bandit problem with discounting." *Theory of Probability & Its Applications* 35 (2): 307–317.

Robbins, Herbert. 1952. "Some aspects of the sequential design of experiments." *Bull. Am. Math. Soc.* 58:527–535.

Rothschild, Michael. 1974. "A two-armed bandit theory of market pricing." *J. Econom. Theory* 9 (2): 185–202.

Schachermayer, W., and W. Schachinger. 1999. "Is there a predictable criterion for mutual singularity of two probability measures on a filtered space?" *Teor. Veroyatnost. i Primenen.* 44 (1): 101–110.

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

Simon, Thomas. 2000. "Support theorem for jump processes." *Stochastic Process. Appl.* 89 (1): 1–30.

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

Veatch, Michael H, and Lawrence M Wein. 1996. "Scheduling a make-to-stock queue: Index policies and hedging points." *Operations Research* 44 (4): 634–647.

Washburn, R. 2008. "Application of multi-armed bandits to sensor management." *Foundations and Applications of Sensor Management:*153–175.

Weber, Richard. 1992. "On the Gittins index for multiarmed bandits." *Ann. Appl. Probab.* 2 (4): 1024–1033.

Weber, R.R., and G. Weiss. 1990. "On an index policy for restless bandits." *Journal of Applied Probability:*637–648.

———. 1991. "Addendum to'On an index policy for restless bandits'." *Advances in Applied probability:*429–430.

Weitzman, Martin L. 1979. "Optimal Search for the Best Alternative." *Econometrica* 47 (3): 641–654.

Whittle, P. 1980. "Multi-armed bandits and the Gittins index." *J. Roy. Statist. Soc. Ser. B* 42 (2): 143–149.

———. 1981. "Arm-Acquiring Bandits." *The Annals of Probability* 9 (2): 284–292.

———. 1988. "Restless bandits: activity allocation in a changing world." *J. Appl. Probab.* No. Special Vol. 25A:287–298.

Wonham, W. M. 1968. "On the separation theorem of stochastic control." *SIAM J. Control* 6:312–326.

5 Appendix A: Technical Appendix

Lemmas 1 through 3 are used to establish theorem 1 and lemmas 4 through 9 to establish theorem 2.

We will say that the discrete setting is in place when we are using the following notation.

Definition 8 (Discrete setting)**.** *For $\delta > 0$ and $i \in \mathbb{N}$ let $t_i = i\delta$. For càdlàg processes P, H, R, etc., we write P_i, H_i, R_i for $P_{t_i}, H_{t_i}, R_{t_i}$. For a càglàd process like U, we write U_i for the right limit U_{t_i+}. The value of a p.o. step control with step size δ can be expressed as*

$$J^{p.o.}(\mathcal{U}) = \mathbb{E}\left(\sum_{i=0}^{\infty} \zeta_i\big(U_i\gamma(\Theta, H_i) + (1-U_i)k\big)\right), \tag{44}$$

where

$$\zeta_i = (1-e^{-\rho\delta})e^{-\rho i\delta}, \quad \gamma(\theta,h) = \frac{1}{1-e^{-\rho\delta}}\,\mathbb{E}\left(\int_0^{\delta} \rho e^{-\rho t} dR_t \middle| (\Theta_0, H_0, R_0) = (\theta, h, 0), U_t \equiv 1\right). \tag{45}$$

Similarly, the value of a se. step control with step size δ is

$$J^{se.}(\mathcal{U}) = \mathbb{E}\left(\sum_{i=0}^{\infty} \zeta_i\big(U_i\overline{\gamma}(P_i, H_i) + (1-U_i)k\big)\right), \tag{46}$$

where

$$\overline{\gamma}(p,h) = p\gamma(1,h) + (1-p)\gamma(0,h). \tag{47}$$

Lemma 1. *Under assumptions 1, and 3, the value functions for the p.o. and se. problem are well-defined and finite.*

Proof. *Step 1.* We claim that the set of controls for the p.o. and se. problem, discretized or not, is not empty. For the p.o. problem, it is sufficient to show that constant control processes F are admissible. This is a consequence of Jacod and Shiryaev (2003, proposition III.2.42) stating that existence and local uniqueness holds for the martingale problem associated to deterministic characteristics. For the se. problem, assumption 3 implies the existence of controls with constant control process.

Step 2. Let $\mathcal{U}$ be a p.o. control and let (B, C, μ) be the characteristics of R. The uniform bound on $\int(|r|^2 \wedge |r|)K(\theta, h, dr)$ in assumption 1 implies that the process $(|r|^2 \wedge |r|) * \mu$ is increasing and of integrable variation. It follows from Jacod and Shiryaev (2003, proposition II.2.29b) that R is a special $(\mathbb{F}, \mathbb{P})$-semimartingale. Therefore, there is a unique predictable process of integrable variation A such that $R - A$ is a local martingale. By the same proposition, A satisfies

$$dA_t = \left(\beta(\Theta, H_t)U_t + k(1-U_t) + U_t \int \big(r - \chi(r)\big)K(\Theta, H_t, dr)\right) dt. \tag{48}$$

The uniform bounds on $\beta(\theta, h)$ and $\int(|r|^2 \wedge |r|)K(\theta, h, dr)$ imply that A_t has at most linear growth

in t. Consequently,

$$J^{\text{p.o.}}(\mathcal{U}) = \mathbb{E}\left(\int_0^\infty \rho e^{-\rho t} dR_t\right) = \mathbb{E}\left(\int_0^\infty \rho e^{-\rho t} dA_t\right) < \infty. \tag{49}$$

The bound on A is uniform in U. Therefore $V^{\text{p.o.}}$ is finite.

Step 3. Now let $\mathcal{U}$ be a se. control. The process P is bounded by assumption 3. It follows by assumption 1 that the integrand in the definition of $J^{\text{se.}}(\mathcal{U})$ in (23) is bounded. Therefore $J^{\text{se.}}(\mathcal{U})$ is finite. The bound is uniform in the control U and therefore $V^{\text{se.}}$ is finite. ■

Lemma 2. *Under assumptions 1 and 2, every p.o. control can be transformed into a se. control with the same value, which implies*

$$V^{p.o.}(p_0, h_0) \leq V^{se.}(p_0, h_0), \quad V^{p.o.,\delta}(p_0, h_0) \leq V^{se.,\delta}(p_0, h_0). \tag{50}$$

Proof. For a p.o. control

$$\big((\Omega, \mathcal{F}, \mathbb{F}, \mathbb{P}), U, \Theta, H, R, p_0, h_0\big) \tag{51}$$

we let P be the unique càdlàg process satisfying $P_t = \mathbb{E}(\Theta|\mathcal{F}_t^R)$. We claim that

$$((\Omega, \mathcal{F}, \mathbb{F}^R, \mathbb{P}), U, P, H, p_0, h_0) \tag{52}$$

is a se. control with the same value as the p.o. one. If the p.o. control is a step control, then the se. control is a step control as well.

Step 1. If $p_0 = 0$ or $p_0 = 1$, then the constant process $P_t = p_0$ yields the desired se. control. Otherwise, the measure $\mathbb{P}$ can be conditioned on the type Θ of the agent. For $\theta \in \{0, 1\}$, this yields measures $\mathbb{P}_\theta$ such

$$\mathbb{P}_\theta(\Theta = \theta) = 1, \quad \mathbb{P} = p_0\mathbb{P}_1 + (1 - p_0)\mathbb{P}_0. \tag{53}$$

The restriction of the measure $\mathbb{P}_\theta$ to $\mathcal{F}^R$ is determined by the martingale problem $\mathcal{I}(\mathcal{F}_0^R, R \mid \delta_0; B, C, \mu)$, where (B, C, μ) and (b, c, ν) are defined by equations (2) and (3) with Θ replaced by θ. By Jacod and Shiryaev (2003, theorem II.2.42), $\mathbb{P}_\theta$ can equivalently be characterized as the solution of the following martingale problem: for all functions $f \in C_b^2(\mathbb{R}^3)$, the process

$$\begin{aligned} M_t^{f,\theta} = f(R_t) - f(R_0) - \int_0^t \frac{\partial f}{\partial r}(R_{s-}) b_s ds - \frac{1}{2}\int_0^t \frac{\partial^2 f}{\partial r^2}(R_{s-}) c_s ds \\ - \int_0^t \int_{\mathbb{R}^n} \left(f(R_{s-} + r) - f(R_{s-}) - \frac{\partial f}{\partial r}(R_{s-})\chi(r)\right)\nu_s(dr)ds \end{aligned} \tag{54}$$

is an $\mathbb{F}^R$-martingale under $\mathbb{P}_\theta$. Moreover, the initial condition $\mathbb{P}_\theta(R_0 = 0) = 1$ holds.

Step 2. We claim that the measures $\mathbb{P}_1$ and $\mathbb{P}_0$ are equivalent on $\mathcal{F}_t^R$ for all t. The local bounds in assumption 2 and the continuity of H give an increasing sequence of stopping times $(S_n)_{n=1,2,\ldots}$

such that the expressions in (18) are bounded. Setting

$$\Lambda_n = \frac{1}{2}\int_0^{S_n} U_t\sigma^2\phi_1(H_t)^2dt + \\ + \int_0^{S_n}\int_{\mathbb{R}} \big(\phi_2(H_t,r)\log\phi_2(H_t,r) - \phi_2(H_t,r) + 1\big)U_tK(0,H_t,dr)dt \quad (55)$$

for $n = 1, 2, \ldots$ we obtain that $\mathbb{E}(e^{\Lambda_n}) < \infty$. When U is constant, then Cheridito, Filipović, and Yor (2005, theorem 2.4) implies the equivalence of $\mathbb{P}_1$ and $\mathbb{P}_0$ on $\mathcal{F}_t^R$ for each t. We will generalize their proof to non-constant strategies, assuming that the bounds on Λ_n are uniform in u and that local uniqueness holds for the martingale problem characterizing $\mathbb{P}_\theta$.

Let R^c be the continuous local martingale part of R relative to $(\mathcal{F}^R, \mathbb{P}_0)$. Moreover, let μ^R denote the integer valued random measure associated to the jumps of R, see Jacod and Shiryaev (2003, II.1.16). Its $(\mathbb{F}^R, \mathbb{P}_0)$-compensator is $U_tK(0, H_t, dr)dt$. (Recall that H is $\mathbb{F}^R$-predictable.) We claim that

$$D = \mathcal{E}(L) \quad \text{with} \quad L = \phi_1(H)\cdot R^c + \big(\phi_2(H,r) - 1\big) * \big(\mu^R - U_tK(0,H_t,dr)dt\big) \quad (56)$$

is the $\mathbb{F}^R$-density process of $\mathbb{P}_1$ with respect to $\mathbb{P}_0$. The stochastic integration in (56) is performed under the measure $\mathbb{P}_0$. Lépingle and Mémin (1978, Théorème IV.3) showed that the condition $\mathbb{E}(e^{\Lambda_n}) < \infty$ implies that D^{S_n} is a non-negative, uniformly integrable $(\mathbb{F}^R, \mathbb{P}_0)$-martingale. Consequently, $D_{S_n}\cdot\mathbb{P}_0$ is a probability measure. Equation (16) implies that

$$(M^{f,1})_t^{S_n} = (M^{f,0})_t^{S_n} - \int_0^t \frac{1}{D_{s-}^{S_n}} d\big\langle (M^{f,0})^{S_n}, D^{S_n}\big\rangle_s.$$

Then Girsanov's theorem in the form of Jacod and Shiryaev (2003, III.3.11) shows that $(M^{f,1})^{S_n}$ is an $(\mathbb{F}^R, D_{S_n}\cdot\mathbb{P}_0)$-martingale. By definition, it is also an $(\mathbb{F}^R, \mathbb{P}_1)$-martingale. Since local uniqueness holds for the martingale problem (54) with initial condition $R_0 = 0$, $\mathbb{P}_1$ is equal to $D_{S_n}\cdot\mathbb{P}_0$ on $\mathcal{F}_{S_n}^R$. Then for each $t \geq 0$ and $A \in \mathcal{F}_t^R$, one has $A \cap \{t < S_n\} \in \mathcal{F}_{S_n\wedge t}^R$ and

$$\mathbb{P}_1(A) = \lim_{n\to\infty}\mathbb{P}_1(A\cap\{t<S_n\}) = \lim_{n\to\infty}\mathbb{E}_{\mathbb{P}_0}(D_{S_n\wedge t}\mathbb{1}_A\mathbb{1}_{\{t<S_n\}}) \\ = \lim_{n\to\infty}\mathbb{E}_{\mathbb{P}_0}(D_t\mathbb{1}_A\mathbb{1}_{\{t<S_n\}}) = \mathbb{E}_{\mathbb{P}_0}(D_t\mathbb{1}_A),$$

which shows that D is the $\mathbb{F}^R$-density process of $\mathbb{P}_1$ relative to $\mathbb{P}_0$. Furthermore, D is strictly positive because it is the Doléan-Dade exponential of a process L with jumps strictly greater than -1, which implies that $\mathbb{P}_1$ and $\mathbb{P}_0$ are locally equivalent. Then also $\mathbb{P}_1$ and $\mathbb{P}$ are locally equivalent and

$$P_t = \mathbb{E}(\Theta|\mathcal{F}_t^R) = \frac{d(\Theta\mathbb{P})|_{\mathcal{F}_t^R}}{d\mathbb{P}|_{\mathcal{F}_t^R}} = \frac{p_0 d\mathbb{P}_1|_{\mathcal{F}_t^R}}{d\mathbb{P}|_{\mathcal{F}_t^R}} = \frac{p_0(d\mathbb{P}_1|_{\mathcal{F}_t^R})/(d\mathbb{P}_0|_{\mathcal{F}_t^R})}{(d\mathbb{P}|_{\mathcal{F}_t^R})/(d\mathbb{P}_0|_{\mathcal{F}_t^R})} = \frac{p_0D_t}{p_0D_t + 1 - p_0} \quad (57)$$

is the $\mathbb{F}^R$-density process of $\mathbb{P}_1$ with respect to $\mathbb{P}$.

Step 3. It remains to calculate the $(\mathbb{F}^R, \mathbb{P})$-characteristics (B, C, μ) of the filter P. Define

$$\psi(x) = \frac{p_0 x}{p_0 x + 1 - p_0}, \quad \psi^{-1}(x) = \frac{1 - p_0}{p_0} \frac{x}{1 - x} \tag{58}$$

such that $P = \psi(\mathcal{E}(L))$ holds. To calculate the jumps of the filter, we write

$$\begin{aligned} P_t = \psi\big(\mathcal{E}(L)_{t-} + \Delta\mathcal{E}(L)_t\big) = \psi\big(\mathcal{E}(L)_{t-} + \mathcal{E}(L)_{t-}\Delta L_t\big) = \psi\big(\psi^{-1}(P_{t-})(1 + \Delta L_t)\big) = \\ = \psi\big(\psi^{-1}(P_{t-})\phi_2(H_t, \Delta R_t)\big) = \frac{P_{t-}\ \phi_2(H_t, \Delta R_t)}{P_{t-}\ \phi_2(H_t, \Delta R_t) + 1 - P_{t-}}. \end{aligned} \tag{59}$$

This implies $\Delta P_t = j(P_{t-}, H_t, \Delta R_t)$ with j defined as in equation (21). By Jacod and Shiryaev (2003, theorem III.3.40), the third $(\mathbb{F}^R, \mathbb{P})$-characteristic of the reward process is $U_t\ \overline{K}(P_t, H_t, dr)dt$. Therefore the third $(\mathbb{F}^R, \mathbb{P})$-characteristic of the filter is

$$\mu(dt, dr) = U_t\ (j_* \overline{K})(P_t, H_t, dr)dt, \tag{60}$$

where the push forward measure $j_*\overline{K}$ is as defined in equation (20).

The continuous local martingale part of the filter with respect to the measure $\mathbb{P}_0$ is

$$\begin{aligned} P^c = \psi(\mathcal{E}(L))^c = \psi'(\mathcal{E}(L)_-) \cdot \mathcal{E}(L)^c = \big(\psi'(\mathcal{E}(L)_-)E(L)_-\big) \cdot L^c = \big(\psi'(\mathcal{E}(L)_-)E(L)_-\phi_1(H)\big) \cdot R^c \\ = \big(\psi' \circ \psi^{-1}(P_-)\psi^{-1}(P_-)\phi_1(H)\big) \cdot R^c = \big(P_-(1 - P_-)\phi_1(H)\big) \cdot R^c. \end{aligned} \tag{61}$$

The quadratic variation $\langle P^c, P^c \rangle$ remains the same when the measure $\mathbb{P}_0$ is replaced by the equivalent measure $\mathbb{P}$, see Jacod and Shiryaev (2003, theorem III.3.11-13). This determines the second characteristic of the filter as follows.

$$dC_t = d\langle P^c, P^c \rangle_t = P_{t-}^2 (1 - P_{t-})^2 \phi_1(H_t)^2 U_t \sigma(H_t)^2 dt \tag{62}$$

Since the filter $P_t = \mathbb{E}(\Theta | \mathcal{F}_t^R)$ is an $(\mathbb{F}^R, \mathbb{P})$-conditional expectation, it is a martingale under $\mathbb{P}$. It follows by Jacod and Shiryaev (2003)[proposition II.2.29] that its first characteristic must satisfy

$$B = -(r - \chi(r)) * \mu. \tag{63}$$

Thus we have shown that the characteristics of the filter are as required in the definition of se. controls.

Step 4. It remains to show that the p.o. and se. strategy have the same value. Taking

$\mathbb{F}^R$-optional projections in equation (49) one obtains that

$$J^{\text{p.o.}}(\mathcal{U}) = \mathbb{E}\bigg(\int_0^\infty \rho e^{-\rho t} \Big(\big(P_t\beta(1,H_t) + (1-P_t)\beta(1,H_t)\big)U_t + k(1-U_t) + \\ + U_t \int \big(r - \chi(r)\big)\big(P_t K(1,H_t,dr) + (1-P_t)K(1,H_t,dr)\big)\Big) dt \bigg), \quad (64)$$

which is equal to $J^{\text{se.}}(\mathcal{U})$ defined in equation (25). ■

Lemma 3. *Under assumption 3, there exists for every se. step control a p.o. step control of at least the same value, which implies the inequality $V^{p.o.,\delta} \geq V^{se.,\delta}$.*

Proof. We work in the discrete setting. To avoid confusion, we will mark objects of the se. problem with a tilde. The se. control problem is that of controlling the discrete time Markov chain $(\widetilde{P}_{t_i}, \widetilde{H}_{t_i})$ with $t_i = i\delta$. Controls are restricted to be $\{0,1\}$-valued. It is well-known that optimal Markov controls exist for such problems, see e.g. Seierstad (2009). We will prove the lemma by showing that every se. step Markov control corresponds to a p.o. step control of the same value.

So we start with an optimal step Markov control $\widetilde{\mathcal{U}}$ and write its control process in the form

$$\widetilde{U}_t = \sum_i f_i(\widetilde{P}_{t_i}, \widetilde{H}_{t_i}) \mathbb{1}_{(t_i, t_{i+1}]}. \quad (65)$$

We will construct the p.o. control on the space $D(\mathbb{R}^3)$ with its natural filtration $\mathbb{F}$, sigma algebra $\mathcal{F}$ and coordinate process $X = (\Theta, H, R)$. For $(u,\theta,h,r) \in \{0,1\}^2 \times \mathbb{R}^2$, let $\mathbb{P}_{u;\theta,h,r}$ be the unique probability measure on $\mathcal{F}$ such that Θ is the constant process $\Theta_t = \theta$, H satisfying (1) and $H_0 = h$, and R has semimartingale characteristics (2)–(3) and satisfies $R_0 = r$. By a straight-forward argument (see e.g. the proof of Jacod and Shiryaev (2003, Corollary III.2.42)), $\mathbb{P}_{u;\theta,h,r}$ depends measurably on (u,θ,h,r). This allows us to inductively define probability measures $\mathbb{P}^n$ and càdlàg processes P^n on $D(\mathbb{R}^3)$ as follows.

$$\mathbb{P}^0 = p_0 \mathbb{P}_{f_0(p_0,h_0);1,h_0,0} + (1-p_0)\mathbb{P}_{f_0(p_0,h_0);0,h_0,0}, \qquad P^0_t = \mathbb{E}_{\mathbb{P}^0}(\Theta_0|\mathcal{F}^R_t), \quad (66)$$

$$\mathbb{P}^n = \mathbb{P}^{n-1} \otimes_{t_n} \mathbb{P}_{f_n(P^{n-1}_{t_n}, H_{t_n});\Theta_{t_n}, H_{t_n}, R_{t_n}}, \qquad P^n_t = \mathbb{E}_{\mathbb{P}^n}(\Theta_0|\mathcal{F}^R_t). \quad (67)$$

Here we have used the following notation: When $\mathbb{P}$ is a probability measure on $D(\mathbb{R}^3)$ and $(\mathbb{Q}_x)_{x\in\mathbb{R}^3}$ is a stochastic kernel from $\mathbb{R}^3$ to $D(\mathbb{R}^3)$, then $\mathbb{P} \otimes_t \mathbb{Q}_X$ is the unique probability measure on $D(\mathbb{R}^3)$ such that the law of the stopped process X^t is equal to $\mathbb{P}$ on $\mathcal{F}_t$ and such that the conditional law of the time-shifted process $(X_{t+s})_{s\geq 0}$ given $X_t = x$ is $\mathbb{Q}_x$. The notation is explained and relevant results are proven in Stroock and Varadhan (2006, 6.1.2, 6.1.3 and 1.2.10) for continuous processes. The relevant results on Skorokhod space are Jacod and Shiryaev (2003, lemmas III.2.43-48), but the notation is not used there.

The measures $\mathbb{P}^n$ and $\mathbb{P}^m$ agree on $\mathcal{F}_{t_n \wedge t_m}$ and the processes P^n and P^m a.s. agree on $[0, t_n \wedge t_m]$. Therefore there is a unique measure $\mathbb{P}$ that coincides with $\mathbb{P}^n$ on $\mathcal{F}_{t_n}$ for all n. Furthermore, there

is a unique càdlàg process P that is a.s. equal to P^n on $[0, t_n]$ for all n. We define

$$U_t = \sum_i f_i(P_{t_i}, H_{t_i}) \mathbb{1}_{(t_i, t_{i+1}]} \tag{68}$$

and claim that

$$\mathcal{U} = \big((D(\mathbb{R}^3), \mathcal{F}, \mathbb{F}, \mathbb{P}), U, \Theta, H, R, p_0, h_0\big) \tag{69}$$

is a p.o. control. So we have to verify that U can be written as a predictable functional of R. (We say that a process can be written as a (predictable) functional of R if it coincides $\mathbb{P}$-a.s. with $F(R)$, where F is an adapted (predictable) process on $D(\mathbb{R})$.) We proceed by induction and claim that for all n, the stopped processes U^{t_n} and H^{t_n} can be written as functionals of R. For $n = 0$ there is nothing to prove. For the inductive step, we observe that $U^{t_{n+1}}$ and $H^{t_{n+1}}$ depend on U^{t_n}, P_{t_n}, and H_{t_n}. These expressions can be written as functionals of R by the inductive assumption and the claim follows. Therefore U can be written as a $U = F(R)$, where F is a predictable process on $D(\mathbb{R})$. F is admissible because it is a step process. The process Θ is a.s. constant and can be identified with a $\{0, 1\}$-valued random variable. H satisfies (1) and R has characteristics (2)-(3) under $\mathbb{P}$. Thus we have verified that (69) is a p.o. control. Then the proof of lemma 2 shows that

$$\widehat{\mathcal{U}} = \big((D(\mathbb{R}^3), \mathcal{F}, \mathbb{F}^R, \mathbb{P}), U, P, H, p_0, h_0\big) \tag{70}$$

is a se. control with the same value. Since U is a step process, assumption 3 implies that (P, H) has the same distribution as $(\widetilde{P}, \widetilde{H})$. It follows that

$$J^{\text{se.}}(\widetilde{\mathcal{U}}) = J^{\text{se.}}(\widehat{\mathcal{U}}) = J^{\text{p.o.}}(\mathcal{U}). \tag{71}$$

■

Lemma 4. *Under assumptions 1, 3, and 4, se. controls can be approximated arbitrarily well by se. step controls, which implies*

$$V^{se.}(p_0, h_0) = \sup_{\delta} V^{se.,\delta}(p_0, h_0) \tag{72}$$

Proof. It is sufficient to show the inequality $\leq$ in (72). The reverse inequality holds by definition.

Step 1. Let $\mathcal{U} \in \mathbb{U}^{\text{se.}}_{p_0,h_0}$ be a se. control with control process U. The first step is to represent the control process by the random measure

$$Q(dt, du) = \big(U_t \delta_1(du) + (1 - U_t)\delta_0(du)\big)dt. \tag{73}$$

Q is a predictable random measure on $\mathbb{R}_{\geq 0} \times \{0, 1\}$ whose marginal on $\mathbb{R}_{\geq 0}$ is the Lebesgue measure. Conversely, for any such random measure there exists a representation as in (73) with a predictable process U, see Jacod and Shiryaev (2003, II.1.7.(i) and I.3.13). Equations (1) and (19) characterize the evolution of the process (P, H) and are equivalent to the following martingale problem: for all

$f \in C_b^2(\mathbb{R}^2)$, the process

$$f(P_t,H_t) - f(P_0,H_0) - \iint_{[0,t]\times\{0,1\}} \Bigg(\frac{\partial f}{\partial h}(P_{s-},H_s)\big(u\alpha(1,H_s) + (1-u)\alpha(0,H_s)\big) + \frac{\partial f}{\partial p}(P_{s-},H_s)b_s + \frac{1}{2}\frac{\partial^2 f}{\partial p^2}(P_{s-},H_s)c_t + \int_{\mathbb{R}^n} \Big(f(P_{s-}+q,H_s) - f(P_{s-},H_s) - \frac{\partial f}{\partial p}(P_{s-},H_s)\chi(q) \Big)\mu_s(dr) \Bigg) Q(ds,du) \quad (74)$$

is a martingale. Moreover, the following initial condition holds:

$$\mathbb{P}(P_0 = p_0) = \mathbb{P}(H_0 = h_0) = 1 \quad (75)$$

Step 2. Let M be the space of all probability measures on $\mathbb{R}_{\geq 0} \times \{0,1\}$ whose marginal on $\mathbb{R}_{\geq 0}$ is the Lebesgue measure. We endow M with the topology of vague convergence, checked on compactly supported continuous functions. This turns M into a compact metric space. Let $\mathcal{M}$ be the Borel sigma algebra on M. We give M the natural filtration $\mathbb{M} = (\mathcal{M}_t)_{t\geq 0}$ generated by $\mathbb{1}_{[0,t]} \cdot q$ for $q \in M$. The canonical space for the se. control problem is $M \times D(\mathbb{R}^2)$. Elements of this space will be denoted by (Q,P,H). We define a *se. control rule* to be a probability measure $\mathbb{P}$ on the space $M \times D(\mathbb{R}^2)$ that solves the martingale problem (74) and satisfies the initial condition (75). The value of a se. control rule is defined as

$$\mathbb{E}\Bigg(\iint_{[0,\infty)\times\{0,1\}} \rho e^{-\rho t} \Big(\overline{\beta}(P_{t-},H_{t-})u + k(1-u) + u\int \big(r - \chi(r)\big)\overline{K}(P_{t-},H_{t-},dr) \Big) Q(dt,du) \Bigg), \quad (76)$$

where the expectation is taken with respect to $\mathbb{P}$. By the above considerations, any se. control in the sense of definition 4 induces a se. control rule and vice versa. Therefore, the se. problem is equivalent to maximizing (76) over all se. control rules.

Step 3. Let $(p_0,h_0) \in \{0,1\} \times \mathbb{R}$ and $Q \in M$. Then Q corresponds to a deterministic control process U. Let $D(\mathbb{R}^2)$ be Skorokhod space with its natural filtration $\mathbb{F}$ and sigma algebra $\mathcal{F}$. By assumption 3, there is one and only one solution measure $\mathbb{S}_Q$ on $\mathcal{F}$ to the martingale problem (74), (75). We claim that $\mathbb{S}_Q$ is continuous in $Q \in M$. This follows from Jacod and Shiryaev (2003, theorem IX.3.39). The verification of the conditions of the theorem is straight-forward and goes along the lines of Jacod and Shiryaev (2003, theorem IX.4.8). The assumptions that are needed are assumption 1 providing uniform bounds on the coefficients and establishing the continuity of α, assumption 3 implying well-posedness of the martingale problem (74) for (P,H) under the deterministic control process U, and assumption 4 establishing the continuity of β, σ, K in an appropriate sense.

Step 4. Let $\mathbb{P}$ be a se. control rule and let Z stand for (P,H). Using disintegration, $\mathbb{P}$ can be written in the form

$$\mathbb{P}(dQ,dZ) = \mathbb{P}^M(dQ)\mathbb{P}_Q^{D(\mathbb{R}^2)}(dZ). \quad (77)$$

Then for $\mathbb{P}^M$-a.e. Q, the measure $\mathbb{P}_Q^{D(\mathbb{R}^2)}$ is equal to the measure $\mathbb{S}_Q$ from step 3. (We have used assumption 3 here.) It is well-known that any measure valued control $Q \in M$ can be approximated in the vague topology by a sequence $\psi^n(Q)$ of measures of the form

$$\psi^n(Q)(dt, du) = \sum_{i=1}^{\infty} \big(U_i^n \delta_1(du) + (1 - U_i^n)\delta_0(du)\big) \mathbb{1}_{(t_i^n, t_{i+1}^n]} dt, \tag{78}$$

where $\psi^n : M \to M$ are $\mathbb{M}$-adapted mappings. This result is known under the name chattering lemma and can be found e.g. in Mazliak (1993, theorem 2.2) or Fleming (1989). Furthermore, t_i^n can be chosen of the form $t_i^n = i\delta^n$ for some sequence of numbers $\delta^n > 0$. Let $\mathbb{P}^{M,n}$ be the push-forward measure of $\mathbb{P}^M$ under ψ^n. Moreover, let

$$\mathbb{P}^n(dQ, dZ) = \mathbb{P}^{M,n}(dQ)\mathbb{S}_Q(dZ). \tag{79}$$

Then each $\mathbb{P}^n$ corresponds to a se. step control $\mathcal{U}^n \in \mathbb{U}_{p_0,h_0}^{\text{se.},\delta^n}$ as explained in step 1. By the result from step 3, $\mathbb{S}_Q$ is continuous in Q. It follows that $\mathbb{P}^n \to \mathbb{P}$ weakly. The value of the control given by the expression in equation (76) is continuous in $\mathbb{P}$ with the weak topology. Therefore $J^{\text{se.}}(\mathcal{U}^n) \to J^{\text{se.}}(\mathcal{U})$. Thus we have shown that the value of a control rule can be approximated arbitrarily well by the value of a step control. ■

Lemmas X through X are used to establish Theorem 2.

Lemma 5. *Under assumptions 1–5, the value function $V(p_0, h_0)$ is convex non-decreasing in p_0 and non-decreasing in h_0. The same statement holds about the discrete time value function $V^\delta(p_0, h_0)$.*

Proof. *Step 1.* Let $\mathcal{U}$ be a se. control with deterministic control process U. Then the martingale property of P can be used to express the value of $\mathcal{U}$ as follows.

$$\begin{aligned} J^{\text{se.}}(\mathcal{U}) = p_0\, \mathbb{E}\left(\int_0^\infty \beta(1, H_t)U_t + k(1 - U_t) + U_t \int \big(r - \chi(r)\big)K(1, H_t, dr)\right) + \\ + (1 - p_0)\, \mathbb{E}\left(\int_0^\infty \beta(0, H_t)U_t + s(1 - U_t) + U_t \int \big(r - \chi(r)\big)K(0, H_t, dr)\right) \end{aligned} \tag{80}$$

This expression is linear in p_0 and non-decreasing in (p_0, h_0) by assumption 5.

Step 2. Now, let $\mathcal{U}$ be a general se. control. As described in the proof of lemma 4, $\mathcal{U}$ corresponds to a measure $\mathbb{P}$ on the canonical space $M \times D(\mathbb{R}^2)$. Elements of this space will be denoted by (Q, Z). Thus Q corresponds to a sample path of the control process U and Z is sample path of (P, H). Using disintegration, $\mathbb{P}$ can be represented as follows.

$$\mathbb{P}(dQ, dZ) = \mathbb{P}^M(dQ)\mathbb{P}_Q^{D(\mathbb{R}^2)}(dZ), \tag{81}$$

where $\mathbb{S}_Q$ is the unique solution to the martingale problem associated to the deterministic control Q, see step 3 in the proof of lemma 4. Let $\mathcal{U}_Q$ be the se. control corresponding to the control rule

$\mathbb{S}_Q$. Then the value of $\mathcal{U}$ can be expressed as

$$J^{\text{se.}}(\mathcal{U}) = \int_M J^{\text{se.}}(\mathcal{U}_Q)\mathbb{P}^M(dQ). \tag{82}$$

Now step 1 implies that this is linear in p_0 and non-decreasing in (p_0, h_0). By taking the supremum over strategies in $\mathbb{U}^{\text{se.},\delta}_{p_0,h_0}$ and $\mathbb{U}^{\text{se.}}_{p_0,h_0}$, respectively, one obtains that $V^{\text{se.}}$ and $V^{\text{se.},\delta}$ have the desired monotonicity and convexity properties. ■

Lemma 6. *Let assumptions 1–5 hold and assume the discrete setting. Then it is optimal to choose the unknown (known) arm initially in the p.o. problem if and only if it is optimal to do so in the se. problem. Furthermore, the unknown arm is uniquely optimal as an initial choice if $\overline{\gamma}(p_0, h_0) > k$.*

Proof. *Step 1.* We first prove the statement about the optimality of the unknown arm under the condition $\overline{\gamma}(p_0, h_0) > k$ for the se. problem. We fix the initial condition $(P_0, H_0) = (p_0, h_0)$ and work in the discretized setting with step size $\delta > 0$. The Bellman equation says that optimal initial choices for the se. problem are characterized by the equation

$$U_0 \in \arg\max_{u\in\{0,1\}} u\overline{\gamma}(p_0, h_0) + (1-u)k + e^{-\rho\delta}\,\mathbb{E}\left(V^{\text{se.},\delta}(P_1, H_1)\middle| U_0 = u, P_0 = p_0, H_0 = h_0\right). \tag{83}$$

Thus the optimal initial choice for the se. problem depends on the sign of the quantity

$$\begin{aligned}\overline{\gamma}(p_0, h_0) - k + e^{-\rho\delta}\,\mathbb{E}\left(V^{\text{se.},\delta}(P_1, H_1)\middle| U_0 = 1, P_0 = p_0, H_0 = h_0\right)& \\ - e^{-\rho\delta}\,\mathbb{E}\left(V^{\text{se.},\delta}(P_1, H_1)\middle| U_0 = 0, P_0 = p_0, H_0 = h_0\right),&\end{aligned} \tag{84}$$

which is the advantage of the unknown arm over the known arm (up to multiplication by a positive constant). Let h_0^+ be the value that H_1 attains after an initial choice of the unknown arm and h_0^- the value after an initial choice of the known arm. By assumption 5, the inequality $h_0^- \le h_0 \le h_0^+$ holds. Furthermore, note that $P_1 = P_0$ under an initial choice of the known arm, as can be seen from the characteristics of P in definition 4. This can be used together with the monotonicity of the value function in h_0 and its convexity in p_0 to show the following estimate:

$$\begin{aligned}&\mathbb{E}\left(V^{\text{se.},\delta}(P_1, H_1)\middle| U_0 = 1, P_0 = p_0, H_0 = h_0\right) - \mathbb{E}\left(V^{\text{se.},\delta}(P_1, H_1)\middle| U_0 = 0, P_0 = p_0, H_0 = h_0\right)\\ &\quad= \mathbb{E}\left(V^{\text{se.},\delta}(P_1, h_0^+)\middle| U_0 = 1, P_0 = p_0, H_0 = h_0\right) - V^{\text{se.},\delta}(p_0, h_0^-)\\ &\quad\ge \mathbb{E}\left(V^{\text{se.},\delta}(P_1, h_0^+)\middle| U_0 = 1, P_0 = p_0, H_0 = h_0\right) - V^{\text{se.},\delta}(p_0, h_0^+) \ge 0\end{aligned} \tag{85}$$

It follows that (84) is strictly positive when $\overline{\gamma}(p_0, h_0) > s$. In this case, the initial choice of the unknown arm is uniquely optimal for the se. problem.

Step 2. Optimal choices for the p.o. problem are characterized by the same equation (83) as for the se. problem, with $V^{\text{se.},\delta}$ replaced by $V^{\text{p.o.},\delta}$. These two value functions agree by theorem 1. It follows that optimal choices for the p.o. and se. problem agree. ■

Lemma 7. *Under assumptions 1–5, there exists for every p.o. step control a p.o. stopping control that is at least as good.*

Proof. We work in the discrete setting with $\delta > 0$.

Step 1. We claim that the lemma is true when the discount sequence is truncated at stage n. The truncated discount sequence is given by

$$\zeta_i = (1 - e^{-\rho\delta})e^{-\rho i\delta}\mathbb{1}_{i\leq n}. \tag{86}$$

More generally, ζ could be a regular discount sequence of finite horizon n, see Berry and Fristedt (1985, definition 5.2.1). We claim that there exists an optimal stopping control, i.e., a control that never switches from known to the unknown arm. We prove the claim by induction on n. For $n = 0$, there is nothing to prove. Now let ζ have horizon $n + 1$. Let $\mathcal{U} \in \mathbb{U}^{\text{p.o.},\delta}_{p_0,h_0}$ be an optimal control rule for the p.o. problem. (This exists because there are optimal se. step controls that can be transformed into optimal p.o. controls by lemma 3.) Let $U = F(R)$ be the control process associated to $\mathcal{U}$. The inductive hypothesis allows one to assume that for $i \geq 1$, U_i never switches from the known to the unknown arm. If U_0 indicates the unknown arm, the proof is complete. Otherwise U has the form

$$U_0 = 0, \qquad U_i = 1 \text{ for } i = 1, \dots, T, \qquad U_i = 0 \text{ for } i = T + 1, \dots$$

The stage T where the strategy changes from the unknown to the known arm is a stopping time. Given that the known arm is chosen initially, the reward obtained at the first stage is deterministic and does not contain any information about the type. Therefore, there is a modification of U that does not depend on the outcome of the first stage. This makes it possible to define a control rule U^* that skips the first action of U. Formally, and in terms of t instead of i, U^* can be defined as

$$U^*_t = F_{t+\delta}\big((R_{0\vee(t-\delta)})_{t\geq 0}\big). \tag{87}$$

It is easy to verify that this is an admissible control process. Let $\mathcal{U}^*$ be the corresponding p.o. control. We claim that $\mathcal{U}^*$ is at least as good as $\mathcal{U}$. Let H and H^* be the human capital processes under the strategies U and U^*, respectively. Furthermore, let $\mathcal{U}^0$ be the control associated to choosing the known arm all the time. The advantages of $\mathcal{U}^*$ and $\mathcal{U}$ over $\mathcal{U}^0$ are

$$J^{\text{p.o.}}(\mathcal{U}^*) - J^{\text{p.o.}}(\mathcal{U}^0) = \mathbb{E}\left(\sum_{i=0}^{T-1}\zeta_i\big(\gamma(\Theta, H^*_i) - k\big)\right) \geq \mathbb{E}\left(\sum_{i=1}^{T}\zeta_{i-1}\big(\gamma(\Theta, H_i) - k\big)\right) \tag{88}$$

$$J^{\text{p.o.}}(\mathcal{U}) - J^{\text{p.o.}}(\mathcal{U}^0) = \mathbb{E}\left(\sum_{i=1}^{T}\zeta_i\big(\gamma(\Theta, H_i) - k\big)\right) \geq 0. \tag{89}$$

The first inequality holds because choosing the known arm decreases ability, see assumption 5. The second inequality holds because the optimal control $\mathcal{U}$ is at least as good as the control designating

the known arm at all stages. The advantage of $\mathcal{U}^*$ over $\mathcal{U}$ is

$$J^{\text{p.o.}}(\mathcal{U}^*) - J^{\text{p.o.}}(\mathcal{U}) \geq \mathbb{E}\left(\sum_{i=1}^{T}(\zeta_{i-1}-\zeta_i)\big(\gamma(\Theta,H_i)-k\big)\right) = \sum_{i=1}^{T}(\zeta_{i-1}-\zeta_i)\underbrace{\mathbb{E}\Big(\mathbb{1}_{i\leq T}\big(\gamma(\Theta,H_i)-k\big)\Big)}_{=:b_i}. \tag{90}$$

The increment of b_i is

$$b_{i+1}-b_i = \mathbb{E}\Big(\mathbb{1}_{i\leq T}\big(\gamma(\Theta,H_{i+1})-\gamma(\Theta,H_i)\big)\Big) + \mathbb{E}\Big(\mathbb{1}_{i=T}\big(k-\gamma(\Theta,H_{i+1})\big)\Big) \tag{91}$$

The first summand on the right-hand side is non-negative for $i=1,2,\ldots$ because H_i increases while the unknown arm is played. By the $\mathcal{F}^R_{i+1}$-measurability of $\mathbb{1}_{i=T}$ and H_{i+1}, the second summand can be written as

$$\mathbb{E}\Big(\mathbb{1}_{i=T}\big(k-\gamma(\Theta,H_{i+1})\big)\Big) = \mathbb{E}\Big(\mathbb{1}_{i=T}\big(k-\overline{\gamma}(P_{i+1},H_{i+1})\big)\Big) = \mathbb{E}\Big(\mathbb{1}_{i=T}\big(k-\overline{\gamma}(P_{T+1},H_{T+1})\big)\Big).$$

Recall that it is optimal under strategy $\mathcal{U}$ to play the known arm at stage $T+1$. Lemma 6 shows that this implies $k\geq\overline{\gamma}(P_{T+1},H_{T+1})$. This proves

$$b_{i+1}\geq b_i \text{ for } i=1,2,\ldots \tag{92}$$

We also have

$$\sum_{i=1}^{\infty}\zeta_i b_i \geq 0 \tag{93}$$

from equation (89). It is shown in Berry and Fristedt (1985, equation (5.2.8)) that (92) and (93) imply

$$J^{\text{p.o.}}(\mathcal{U}^*) - J^{\text{p.o.}}(\mathcal{U}) = \sum_{i=1}^{\infty}(\zeta_{i-1}-\zeta_i)b_i \geq 0 \tag{94}$$

when ζ is regular. In our case, ζ is regular because it is a truncated geometric discount sequence. Thus we have constructed a strategy $\mathcal{U}^*$ that is at least as good as the optimal strategy $\mathcal{U}$ and never switches from the known to the unknown arm.

Step 2. To drop the assumption that the discount sequence has a finite horizon, one approximates an arbitrary geometric (or, more generally, regular) discount sequence ζ by truncated discount sequences ζ^n with finite horizon. The argument can be found in the proof of Berry and Fristedt (1985, theorem 5.2.2). ■

Lemma 8. *Under assumptions 1–5, the stopping time* $T^*=\inf\{t: V(P_t,H_t)\leq k\}$ *is optimal for the se. problem.*

Proof. Let (P,H) be the process governed by the constant control $U_t=1$ in the sense of definition 4. This is a Feller process by assumptions 1, 3, 4 and Jacod and Shiryaev (2003, theorem IX.4.39). Let $(\widetilde{P},\widetilde{H})$ be the killed version of (P,H) with killing rate ρ, see Peskir and Shiryaev (2006, section

II.5.4). Then we define

$$f(p,h) = \rho\Big(\overline{\beta}(p,h) + \int \big(r - \chi(r)\big)\overline{K}(p,h,dr) - k\Big), \quad f(\partial) = 0, \tag{95}$$

where ∂ denotes the "cemetery point" of the killed process. Moreover, we define

$$I_t = i + \int_0^t f(\widetilde{P}_t, \widetilde{H}_t)dt \tag{96}$$

and $X = (\widetilde{P}, \widetilde{H}, I)$. Then X is a Feller process on the state space

$$E = (\mathbb{R}^2 \cup \{\partial\}) \times \mathbb{R}. \tag{97}$$

Let $(\mathbb{P}_x)_{x\in E}$ denote the family of laws of X starting from the initial condition $X_0 = x$. We associate to it the following family of optimal stopping problems:

$$W(x) = \sup_T \mathbb{E}_x(I_T), \tag{98}$$

where the supremum is over the set of all $\mathbb{F}^X$-stopping times. Since the value function V is attained as the supremum over stopping controls by lemma 7, one obtains that

$$\begin{aligned} W(p,h,i) &= \sup_T \mathbb{E}_{p,h,i}(I_T) = \sup_T \mathbb{E}_{p,h,0}(I_T) + i \\ &= \sup_T \mathbb{E}_{p,h,0}\left(\int_0^T \rho e^{-\rho t}\Big(\overline{\beta}(P_t,H_t) + \int \big(r-\chi(r)\big)\overline{K}(P_t,H_t,dr) - k\Big)dt\right) + i \\ &= V(p,h) - k + i \end{aligned} \tag{99}$$

for $(p,h,i) \in \mathbb{R}^3$. We define the stopping set D as in Peskir and Shiryaev (2006, equation (2.2.5)) by

$$D = \{(\widetilde{p},\widetilde{h},i) \in E : W(\widetilde{p},\widetilde{h},i) \le i\} = \big(\{(p,h) \in \mathbb{R}^2 : V(p,h) \le k\} \cup \{\partial\}\big) \times \mathbb{R}. \tag{100}$$

The equality in (100) holds because $W(\partial, i) = i$, which is obvious from the definitions. The function W is lower semi-continuous because $(\widetilde{P}, \widetilde{H}, I)$ is Feller, see Peskir and Shiryaev (2006, equation (2.2.80)). Therefore the set D is closed. Then the right-continuity of the filtration implies that

$$T^* = \inf\{t \ge 0 : X_t \in D\} = \inf\{t : V(P_t,H_t) \le k\} \tag{101}$$

is a stopping time. Note that $\partial \in D$, which implies $\mathbb{P}(T^* < \infty) = 1$. Then Peskir and Shiryaev (2006, Corollary 2.9) implies that T^* is optimal. ■

Lemma 9. *Under assumption 3, there exists for each se. stopping control a p.o. stopping control with the same value.*

Proof. The idea of the proof is very similar to that of lemma 3. To avoid confusion, we will mark

objects of the se. problem with a tilde. Let $\widetilde{\mathcal{U}}$ be a se. stopping control with control process $\widetilde{U}$. By working on the canonical space $M \times D(\mathbb{R}^2)$ for the se. problem (see the proof of lemma 4), we can assume that the control process is $\mathbb{F}^{\widetilde{P},\widetilde{H}}$-predictable. Then up to a.s. equivalence, it can be represented as $\widetilde{U} = \mathbb{1}_{[\![0,S(\widetilde{P},\widetilde{H})]\!]}(t)$, where S is a stopping time on $D(\mathbb{R}^2)$.

Let $(D(\mathbb{R}^3), \mathcal{F}, \mathbb{F})$ be the canonical path space for $X = (\Theta, H, R)$. As in lemma 3, we let $\mathbb{P}_{u,x}$ denote the law of X under the constant control $U_t = u$ with initial condition $X_0 = x$. Let

$$\mathbb{Q} = p_0 \mathbb{P}_{1,(1,h_0,r_0)} + (1-p_0)\mathbb{P}_{1,(0,h_0,r_0)} \tag{102}$$

and let Q be the unique càdlàg process satisfying $Q_t = \mathbb{E}_{\mathbb{Q}}(\Theta|\mathcal{F}_t^R)$. Since the process H is deterministic under $\mathbb{Q}$, the stopping time $S(Q,H)$ is actually an $\mathbb{F}^R$-stopping time on $D(\mathbb{R}^3)$. It follows that up to a.s. equivalence, it can be written as $T(R)$, where T is a stopping time on $D(\mathbb{R})$. Let

$$\mathbb{P} = \mathbb{Q} \otimes_{T(R)} \mathbb{P}_{0,X_{T(R)}} \tag{103}$$

be the unique probability measure on $D(\mathbb{R}^3)$ such that the law of the stopped process $X^{T(R)}$ is equal to $\mathbb{Q}$ on $\mathcal{F}_{T(R)}$ and such that the conditional law of the time-shifted process $(X_{T(R)+s})_{s\geq 0}$ given $X_{T(R)} = x$ is $\mathbb{P}_{0,x}$. Since the process $\mathbb{1}_{[\![0,T(R)]\!]}$ is piecewise constant, it follows that it is admissible in the sense of definition 1. Thus $\mathbb{P}$ defines a p.o. control

$$\mathcal{U} = \big((D(\mathbb{R}^3, \mathcal{F}, \mathbb{F}, \mathbb{P}), \mathbb{1}_{[\![0,T(R)]\!]}, \Theta, H, R, p_0, h_0)\big). \tag{104}$$

Lemma 2 applied to this control shows that the characteristics of P are as required in definition 4, which means that

$$\widehat{\mathcal{U}} = \big((D(\mathbb{R}^3, \mathcal{F}, \mathbb{F}, \mathbb{P}), \mathbb{1}_{[\![0,T(R)]\!]}, P, H, p_0, h_0)\big) \tag{105}$$

is a separated control with the same value as $\mathcal{U}$. By the local uniqueness assumption 3, (P,H) is equal in law to $(\widetilde{P},\widetilde{H})$. Therefore $\mathcal{U}$ is a p.o. stopping control with the same value as $\widehat{\mathcal{U}}$ and $\widetilde{\mathcal{U}}$. ■