# Robust Experimentation in the Continuous Time Bandit Problem

Farzad Pourbabaee*

Department of Economics, University of California Berkeley

**Abstract**

We consider the experimentation dynamics of a decision maker (DM) in a two-armed bandit setup, where the agent holds *ambiguous* beliefs regarding the distribution of the return process of one arm and is certain about the other one. The DM entertains *Multiplier preferences* à la Hansen and Sargent [2001], thus we frame the decision making environment as a two-player differential game against *nature* in continuous time. We characterize the DM's value function and her optimal experimentation strategy that turns out to follow a cut-off rule with respect to her belief process. The belief threshold for exploring the ambiguous arm is found in closed form and is shown to be increasing with respect to the ambiguity aversion index. We then study the effect of provision of an unambiguous information source about the ambiguous arm. Interestingly, we show that the exploration threshold rises unambiguously as a result of this new information source, thereby leading to more *conservatism*. This analysis also sheds light on the efficient time to reach for an expert opinion.

*JEL classification:* C44; C61; C73; D81; D83

*Keywords:* Model uncertainty; Dynamic experimentation; Variational preferences; Information valuation; Ambiguous diffusion

*Email: farzad@berkeley.edu

414 Evans Hall, University of California, Berkeley, CA 94720

I would like to thank Robert Anderson, Philipp Strack, Gustavo Manso and Demian Pouzo for the support and guidance over the course of this paper, and I am grateful to Haluk Ergin, Chris Shannon and David Ahn for the valuable comments and suggestions. All remaining errors are mine.

# 1 Introduction

There are natural cases where the experimentation shall be performed in ambiguous environments, where the distribution of future shocks is unknown. For example, consider a diagnostician who has two treatments for a particular set of symptoms. One is the conventional treatment that has been widely tested and has a known success rate. Alternatively, there is a second treatment that is recently discovered and is due to further study. The diagnostician shall perform a sequence of experiments on patients to figure out the success/failure rate of the new treatment. However, the adversarial effects of the mistreatment on certain types of patients are fatal, thus the medics must consider the *worst-case* scenario on the patients while evaluating the new treatment. As another case, consider the R&D example of Weitzman [1979], where the research department of an organization is assigned with the task of selecting one of the two technologies producing the same commodity. The research division holds a prior on the generated saving of each technology, but the observations of each alternative during the experimentation stage is obfuscated by ambiguous sources such as the quality of researchers and managerial biases toward one choice. Therefore, the technology that is selected and sent to the development stage must be robust against these sources, because once developed it will be then used in mass production, thus even minor miscalculations in the research stage can lead to huge losses in the sales stage relative to what could have been possibly achieved.

At the core of our paper is an experimentation process between two projects framed as a two-armed bandit problem. The return rate to one arm is known to be $r$, whereas the return rate of the second arm is a binary random variable $\theta \in \{\underline{\theta}, \overline{\theta}\}$, such that $\underline{\theta} \leq r \leq \overline{\theta}$. The decision maker (henceforth DM) holds an initial prior $p_0 = \mathsf{P}\left[\theta = \overline{\theta}\right]$, that can be updated when she invests in the second project and learns its output. At the outset, she has to sequentially choose arms to learn about the unknown return rate while maximizing her net experimentation payoff. Specifically in our model, the observations of the second arm are obfuscated by Wiener process whose distribution, from the perspective of the DM, is unknown and therefore is called the *ambiguous* arm. Central to the agent's decision making problem is her preference for robustness against a candidate set of future shocks' distribution which are concealing the ambiguous arm's return rate. Our investigation of the multiplicity of shocks' distribution is motivated both from the subjective and objective perspectives. Subjectively, the DM might be ambiguity averse and the multiple prior set (for the shock distribution) would be part of her axiomatic utility representation (Gilboa and Schmeidler [1989]). Alternatively, the DM might be subject to an experimentation setup where the results are objectively drawn from a family of distributions, and she wants to maintain a form of robustness against this multiplicity; this is along the lines of *model-uncertainty*

pioneered by Hansen and Sargent [2001] and Hansen et al. [2006].

## 1.1 Summary of results

We frame the decision making environment in which the DM has Multiplier preference, à la Hansen and Sargent [2001], as a two-player continuous time differential game against nature — second player. The DM's goal is to find an allocation strategy between two arms that maximizes her payoff under the distribution picked by the nature. We express the (first player's) payoff function with respect to two control processes: (i) DM's allocation choice process between the two arms, and (ii) the nature's adversarial choice of underlying distribution. The DM follows the *max-min* strategy, namely at every point in time she chooses her allocation weights between two arms, and then the nature picks the shock distribution that minimizes the DM's continuation payoff. We then characterize the value function (to the DM) as a solution to a certain HJBI (Hamilton-Jacobi-Bellman-Isaac) equation.

In this game, the nature's move, i.e choice of the shock distribution, would have two important impacts (with opposite forces) on the DM. First, it affects the current flow payoff of experimentation, and secondly it distorts the DM's posterior formation and consequently her continuation strategy. In the equilibrium the DM knows the nature's best-response strategy, therefore, when she Bayes-updates her belief about $\theta$, she is no longer concerned about *all* possible distributions of shocks. This gives rise to a unique law of motion for the posterior process, and reduces the HJBI equation to a second order HJB equation.

We derive a closed-form expression for the DM's value function with respect to her posterior, i.e $v(p)$, and characterize her robust optimal experimentation strategy. It turns out in the equilibrium her strategy follows a cut-off rule with respect to her belief. Specifically, she switches to the safe arm from the ambiguous arm whenever her posterior drops below a certain threshold $\bar{p}$. We also find a closed-form equation for the cut-off value that allows us to perform a number of comparative statics. In particular, the threshold for selecting the ambiguous arm unambiguously rises as the DM's ambiguity aversion index increases.[1] Also, we establish that the marginal value of receiving *good news* about $\theta$ is increasing, namely $v''(p) \geq 0$.

We then explore the effect of an additional unambiguous information source. In particular, we are interested to know what happens when for e.g the experimentation unit hires an expert to release risky but unambiguous information about $\theta$. The new value function $\tilde{v}(p)$ is obtained in closed-form, and the DM's optimal strategy again turns out to follow a cut-off rule (with a different threshold $\tilde{p}$). Interestingly, we show that under any circumstances, compared to the previous case the value of cut-off rises as a result of the extra information,

[1] The direction of such a response is intuitive, however, the sharp characterization of the threshold via the means of continuous time techniques provides us with the extent of this response.

i.e $\tilde{p} \geq \bar{p}$. Therefore, it is interpreted as though the DM becomes more conservative against choosing the second arm when offered with such information. Lastly, we show the surplus $\tilde{v}(p) - v(p)$ generated by the expert attains its maximum at the range of beliefs where the experimentation unit would otherwise select the ambiguous arm but do not have strong enough feeling and evidence in favor of this decision. Therefore, our model sheds light on the time that is best to reach an expert opinion.

## 1.2 Related literature and organization of the paper

The literature on robust bandit problem is limited, but recently there have been some attempts to bring several aspects of robustness into play. Specifically in the works done by Caro and Gupta [2013] and Kim and Lim [2015] the discrete-time multi-armed bandit problem is studied while the state transition probabilities are drawn from an *ambiguous* set of conditional distributions. In Caro and Gupta [2013] the set of multiple transition probabilities at every period is constrained through a relative entropy condition, whereas Kim and Lim [2015] chooses to impose an entropic penalty cost directly in the objective function of the DM rather than hard thresholding it as a constraint. In a different work Li [2019] studies the multi-armed bandit in which the DM entertains max-min utility and follows a prior-by-prior Bayes updating from her initial *rectangular* multiple prior set, where each candidate distribution in this set is identified by the i.i.d shocks it generates in the future. Our work is different from these treatments in the following aspects: (i) contrary to the first two works the Brownian diffusion treatment of the Markov transitions allows for a richer set of perturbations around benchmark model which extends the scope of robustness that the DM demands; (ii) the continuous time framework lets us to obtain sharp and closed form results on the value function and the optimal experimentation policy that in turn renders the comparative static with regard to parameters of the model and importantly the ambiguity aversion index; (iii) we are explicit about the state variable in our setup, and specifically we characterize it as the DM's posterior process regarding the second arm's return rate; (iv) our setup is flexible enough that can address distinct informational environments such as the effect of the provision of an expert opinion.

In the economic literature, after the seminal work of Gittins [1979], the continuous time problem of optimal experimentation in a noisy environment, where the payoff to the unexplored arm[2] is subject to a Brownian motion is studied in Bolton and Harris [1999] and Keller and Rady [1999]. Aside from these works, there is a growing literature on experimentation in a multiple agent environment where the free-riding issues arise.[3]

Our treatment of robust preferences in continuous time relies heavily on the fundamental

[2]Often the second arm is referred as the *unexplored* one.

[3]A nonexhaustive list includes Keller et al. [2005], Heidhues et al. [2015] and Bonatti and Hörner [2017].

works by Hansen and Sargent [2001], Hansen et al. [2006] and Hansen and Sargent [2011].[4] Our paper is also related to the literature studying the effects of robustness and ambiguity in different decision making frameworks such as Riedel [2009], Cheng and Riedel [2013] and Miao and Rivera [2016]. Also, it is related to the relatively understudied topic of learning under ambiguity.[5] Finally in a set of experimental works with adopting different notions of ambiguity aversion, it has been tested that the ambiguous arm of the experiment has a lower Gittins index that prompts the DM to undervalue the information from exploration. To name a few we can point to Anderson [2012] and Meyer and Shi [1995] in the context of airline choice and Viefers [2012] in the investment choice.

The remainder of the paper is organized as follows. To build intuition, in section 2 we present some of the forces behind the model in a two-period example. Next, in section 3 the full features of experimentation setup and payoff function are explained in a continuous time framework. In section 4, we apply the dynamic programming analysis and present variational characterizations of the value function. Section 5 offers the closed-form expression for value function, properties of the optimal experimentation strategy, and some comparative static results. In section 6, we extend our setup to capture the effect of an additional unambiguous information source. The concluding remarks are presented in section 7 and finally the proofs of all results are expressed in the appendix A.

## 2 Two-period example

Our goal in this example is to highlight the main trade-offs that the DM and her opponent *nature* face in their dynamic interaction. Let $t \in \{1, 2\}$ and at each period the DM allocates her resources between two available choices, namely the safe and the ambiguous project. The time $t$ incremental returns to each arm when she allocates $\mu_t \in [0, 1]$ of her resources to the safe (first) arm and $1 - \mu_t$ to the ambiguous (second) arm are

$$\begin{aligned} \Delta y_{1,t} &= (1 - \mu_t) r \\ \Delta y_{2,t} &= \mu_t \theta + \sqrt{\mu_t} \varepsilon_t. \end{aligned} \tag{2.1}$$

In that $r = 1$ is the return rate of the safe project, and $\theta \in \{0, 2\}$ is the unknown return to the second arm. The DM's prior on this set at period one is given by $p_1 = \mathsf{P}[\theta = 2]$, which is not subject to any ambiguity. However, at each period the return to the second arm is obfuscated by an independent[6] Gaussian shock that could possibly be drawn from two

[4] In a closely related discrete-time framework Epstein and Schneider [2003] and Maccheroni et al. [2006b] present recursive utility representation aimed to capture the preference for robustness.

[5] For e.g. see Marinacci [2002], Epstein and Schneider [2007] and Epstein and Ji [2017].

[6] For simplicity assume $\varepsilon_1$, $\varepsilon_2$ and the period one belief on $\theta$ are independent from each other.

distributions, namely for each $t$ the law of $\varepsilon_t$ belongs to the set $\{\mathcal{N}(-0.5,1),\mathcal{N}(0.5,1)\}$.[7] We take no stance on whether this multiple prior set is the subjective belief of the DM or literally the objective moves that nature takes against the DM. Our solution concept for both cases is the the so-called *max-min*. However, the first situation reflects a decision theoretic choice of an ambiguity averse agent with a subjective multiple prior set, whereas the second interpretation is more in line with the notion of robust decision making.

The timing of this example is as follows. At the beginning of period one DM chooses $\mu_1$. Then, nature *responds* by picking $h_1 \in \{-0.5, 0.5\}$ as the mean of $\varepsilon_1$. The returns to both arms, i.e $\{\Delta y_{1,1}, \Delta y_{2,1}\}$ are realized. DM forms the family of beliefs $\{p_2^{h_1} : h_1\}$ at the beginning of period two, and takes the *appropriate* action $\mu_2$. The nature chooses $h_2$ as the mean of second period's shock. Subsequently the game ends and second period's returns are realized.

What happens at the sub-game perfect equilibrium of this game? For this we need to look at the sub-game starting at $t = 2$. Regardless of DM's action $\mu_2$, the nature always picks $h_2 = -0.5$, because the game ends at this period and $h_2 = -0.5$ is the worst case distribution from the DM's perspective. Because of this triviality of the nature's choice at period two, we drop the index one from $h_1$ and henceforth denote it by $h$, which is the only non-trivial choice of the nature in this example. The DM's posterior beliefs after the realizations of first period returns are

$$p_2^h = \left(1 + \frac{1-p_1}{p_1} \exp\left\{2\left(\sqrt{\mu_1}h + \mu_1 - \Delta y_{2,1}\right)\right\}\right)^{-1} 1_{\{\mu_1>0\}} + p_1 1_{\{\mu_1=0\}}, \quad h \in \{-0.5, 0.5\}. \tag{2.2}$$

It is important to note that the posterior probability is no longer unique, and DM faces a set of posteriors for each choice of nature in period one. Even though that we face a two-player game where the nature's actions are not observable to the DM, but at the equilibrium DM knows the *minimizing* choice of the nature, thereby her family of posteriors effectively reduces to a single posterior induced by the worst case action of the nature say $h^*$. This point becomes more clear as we proceed through the equilibrium analysis. For every member $p_2$ of the posterior set, the DM's optimal action at period two (anticipating that nature will choose $h_2 = -0.5$) is $\mu_2(p_2) = 1_{\{2p_2-0.5>1\}}$, that leads to the expected payoff of $v_2(p_2) = \max\{1, 2p_2 - 0.5\}$. Note that this expectation is with respect to the equilibrium distribution choice of the nature that is $h_2 = -0.5$. Assume the experimenter's intertemporal discount rate is $\delta \in (0, 1]$. Further, let $\mathsf{P}^h$ denote the probability measure induced by the independent product of $\varepsilon_1 \sim \mathcal{N}(h, 1)$ and $\theta \sim p_1$. Therefore, the DM's value function as of beginning of

[7] This set clearly doesn't satisfy the rectangularity condition nor the convexity property of Gilboa and Schmeidler [1989], however it serves only for expositional purposes.

period one is

$$v_1(p_1) = \max_{\mu_1 \in [0,1]} \min_{h \in \{-0.5, 0.5\}} \left\{ \left[ (1-\mu_1) + 2\mu_1 p_1 + \sqrt{\mu_1} h) \right] + \delta \mathsf{E}^h \left[ v_2(p_2^h) \right] \right\}. \tag{2.3}$$

Below we point out to some of the underlying equilibrium forces that will show up in this two period example.

(i) The nature's first period action, or alternatively, the most pessimistic perception of the DM in regard to shock distribution $\varepsilon_1^h$, plays two roles. **Current payoff channel**, in that the nature's choice of $h$ affects the current payoff of the DM by changing the mean return of the ambiguous arm, i.e $\left[ (1-\mu_1) + 2\mu_1 p_1 + \sqrt{\mu_1} h) \right]$. In particular, this is a *positive* force, as higher $h$'s correspond to higher mean flow payoff. **Informational channel**, where the shock distribution $\varepsilon_1^h$ affects the next period belief of the DM, hence changes her course of action and thereby the continuation payoff. This has a *negative* effect, because as $h$ increases, the distribution of $\Delta y_{2,1}$ shifts to the right in the FOSD sense and for a fixed $\Delta y_{2,1}$ lowers the likelihood ratio in (2.2) that in turns depresses the continuation payoff $\mathsf{E}^h \left[ v_2(p_2^h) \right]$. At the equilibrium, nature counteracts these forces and picks the one that its negative effect outweighs the positive one, and thus reduces the DM's payoff more. However, it can not completely balance out the marginal impact of these forces, mainly because we assumed the multiple prior set consists of only two distributions. When the complete mode is laid out in section 3, we allow for quite general multiple prior set, thus nature can precisely cancel out the marginal effects, thereby lowering the DM's payoff as much as possible.

(ii) From the point of view of the DM, there is an option value of experimentation. Specifically, in the first period she selects the ambiguous arm (even partially $0 < \mu < 1$) only to observe the payout of second arm, and then may decide to abandon the ambiguous project depending on the outcome of the first period. In this example, the DM switches back to the safe arm in the second period if her posterior in the equilibrium, i.e $p_2^{h^*}$, drops below a certain threshold, which in this case is 0.75.

(iii) The DM's value function is unambiguously increasing in her initial belief $p_1$ (as can be confirmed from (2.3)), but the marginal value of good news need not be increasing (meaning $v''$ is not always positive). This is mainly due to the finite-horizon setup of the two-period model, which is relaxed in later sections.

(iv) The value function in (2.3) refers to the max-min value of the game, which is associated to the strategic order of actions in which the DM takes her action first and then the nature responds in every period. This is the same approach that we pursue when we

present the complete model. However, one might wonder when does this max-min value coincide with the min-max one? Or in the other words, when does the strategic order of players' actions become irrelevant? In this example the max-min value is strictly less than min-max. Although not related to the study of this paper, but we confirm that with compact and convex action spaces of both players, the von-Neumann minimax theorem could be applied and therefore one can conceive the unique value of the zero-sum game between DM and the nature.

We do not intend to delve deeper into this example and express more specific results and comparative statics, mainly because such analysis will be carried out for the complete model later in the paper.

# 3 Experimentation model

Time horizon is infinite and $t \in \mathbb{R}_+$. There are two projects available to experiment by the DM. Her choice at time $t$ is thus to allocate her resources between two alternatives, namely $\mu_t$ to the ambiguous arm and $1-\mu_t$ to the safe arm. The return process of the projects are[8]

$$\begin{aligned} \mathrm{d}y_{1,t} &= (1-\mu_t)r\mathrm{d}t \\ \mathrm{d}y_{2,t} &= \mu_t\theta\mathrm{d}t + \sigma\sqrt{\mu_t}\mathrm{d}B_t. \end{aligned} \tag{3.1}$$

Here $B$ is a Brownian motion relative to some underlying stochastic basis[9], that represents the shock process, and $\theta$ is unknown to the DM but belongs to the binary set $\{\bar{\theta}, \underline{\theta}\}$, where $\underline{\theta} \leq r \leq \bar{\theta}$. The DM has an initial belief $p_0 = \mathsf{P}\left[\theta = \bar{\theta}\right]$ about $\theta$ which is independent from $B$. The form of return processes in (3.1) follows Bolton and Harris [1999], but we let the DM to associate multiple distributions to the shock process. Specifically, the DM holds a single belief over $\theta$ — so that this represents the uncertainty due to *risk* — but has multiple beliefs regarding the shock distribution $B$ — so this represents the uncertainty due to *ambiguity*.[10]

## 3.1 A framework for modelling ambiguity

Our take of ambiguity or model uncertainty is similar to Hansen et al. [2006] and Hansen and Sargent [2011]. In particular, we assume there is a family of pairs $\left\{(\mathsf{P}^h, B^h) : h \in \mathcal{H}\right\}$

[8] The goal of this section is to study the interplay between ambiguity regarding the new arm and optimal experimentation, thus for simplicity we assume that the conventional arm has a sure return rate of $r$ and is not subject to any source of randomness. Therefore, it is only the second arm that carries the Brownian motion term.

[9] The description of the underlying stochastic basis and the joint structure of processes are explained in the subsection devoted to the *weak formulation*.

[10] This type of uncertainty is sometimes referred to as *model uncertainty* in the literature.

such that for each $h \in \mathcal{H}$, $B^h$ is a Brownian motion under $\mathsf{P}^h$, and DM views this as her multiple prior set. We think of $\mathcal{H}$ – which thus far has not been defined – as the nature's action space, and each $h \in \mathcal{H}$ is deemed as a possible nature's move. We assume there exists a *benchmark* probability specification $\mathsf{P}$ that is *equivalent* (mutually absolutely continuous with respect) to each member of $\mathcal{P} := \{\mathsf{P}^h : h \in \mathcal{H}\}$. The benchmark measure $\mathsf{P}$ and the set $\mathcal{P}$ are interpreted differently based on the context. For example, DM might believe that $\mathsf{P}$ is the underlying probability measure, but considers $\mathcal{P}$ as the approximations of the true distribution because she has preference for robustness. Alternatively, $\mathcal{P}$ could be conceived as the multiple prior set for the ambiguity averse DM in the axiomatic treatment of Gilboa and Schmeidler [1989].

DM has *Multiplier preference* and maximizes the following payoff over an *admissible* set of experimentation strategies $\mathcal{U}$ — with some technical considerations that are elaborated later in the paper:

$$\inf_{h\in\mathcal{H}} \left\{ \mathsf{E}^{\mathsf{P}^h} \left[ \delta \int_0^\infty e^{-\delta t} \mathrm{d}\left(y_{1,t} + y_{2,t}\right) \right] + \alpha H\left(\mathsf{P}^h; \mathsf{P}\right) \right\} \tag{3.2}$$

Here $\delta$ is the time discount rate. The first term in the DM's utility is simply the expected discounted payoff from both projects taken with respect to the measure $\mathsf{P}^h$, and the second term penalizes the belief misspecification using the relative discounted entropy to measure the discrepancy between $\mathsf{P}$ and $\mathsf{P}^h$. Parameter $\alpha$ captures the extent of this penalization, where its larger values associate to smaller penalty. We shall also interpret $\alpha$ as the inverse of ambiguity aversion and relate (3.2) to the *dynamic variational utility representation* of Maccheroni et al. [2006a] and Maccheroni et al. [2006b]. A large $\alpha$ means that the DM does not suffer a lot from ambiguity aversion. In contrast as $\alpha \to 0$, the DM experiences larger utility loss due to severe penalization.

In the next subsection we use the *weak-formulation* approach from the theory of stochastic processes to elaborate and simplify DM's utility function (3.2).

## 3.2 Weak formulation

In this part we present a sound foundation for the joint structure of all the stochastic processes in the model[11]. Let $\left(\Omega, \mathcal{F} = \mathcal{F}_\infty, \mathbf{F} = \{\mathcal{F}_t\}_{t\in\mathbb{R}_+}, \mathsf{P}\right)$ be the stochastic basis, where the filtration satisfies the *usual conditions*.[12] The average rate of return to the ambiguous project $\theta$ is a binary $\mathcal{F}_0$-measurable random variable.

[11] The materials in this subsection might look somewhat technical and unnecessary to some readers, but are essential for rigorous development of the model.

[12] It is right-continuous and $\mathsf{P}$-complete.

**Definition 3.1** (Strategy spaces)**.** The DM's strategy space $\mathcal{U}$ — with a representative point $\mu \in \mathcal{U}$ — is the set of all $\mathbf{F}$-progressive processes[13] taking value in $[0,1]$. The nature's strategy space $\mathcal{H}$ — with a representative point $h \in \mathcal{H}$ — is the space of all *bounded* $\mathbf{F}$-progressive processes.

**Definition 3.2** (Integral forms)**.** For any pair of processes $\{f, g\}$ where $f$ is $g$-integrable[14] we use the alternative notation for integration: $(f \cdot g)_t := \int_0^t f_s \mathrm{d}g_s$. Further, the symbol $\imath$ refers to identity mapping $t \mapsto t$ on $\mathbb{R}_+$. Then the differential return expressions in (3.1) can be represented in the integral form $y_1 = (1-\mu)r \cdot \imath$ and $y_2 = \mu\theta \cdot \imath + \sigma\sqrt{\mu} \cdot B$.

To model the ambiguity we appeal to the weak formulation. In particular, we think of ambiguity as the source that changes the distribution of return process $\{y_1, y_2\}$, but not its sample paths. For this on every finite interval $[0,T]$ we define the probability measure $\mathsf{P}^h_T$ with the following Radon-Nikodym derivative process:

$$\left.\frac{\mathrm{d}\mathsf{P}^h_T}{\mathrm{d}\mathsf{P}}\right|_{\mathcal{F}_t} := L^h_{t,T} = \exp\left\{(h \cdot B)_t - \frac{1}{2}\left(h^2 \cdot \imath\right)_t\right\}, \quad \forall t \leq T \tag{3.3}$$

This relation explains how nature with its choice of $h \in \mathcal{H}$ could induce a new probability measure. The Girsanov's theorem implies that $\mathsf{P}^h_T$ is mutually absolutely continuous with respect to $\mathsf{P}$ — that is often called *equivalent* measure and denoted by $\mathsf{P}^h_T \sim \mathsf{P}$ on $\mathcal{F}_T$. It also implies that the *mean-shifted* process $B^h := B - (h \cdot \imath)$ is a $\mathbf{F}$-Brownian motion under $\mathsf{P}^h_T$ over the interval $[0,T]$. The main catch here is that we can only characterize the perturbations of benchmark probability model $\mathsf{P}$ over finite intervals, that is for example we know how $\mathsf{P}^h$ looks like on $\mathcal{F}_T$ for any finite $T$. However, what is needed for the utility representation in (3.2) is a specification of $\mathsf{P}^h$ on the terminal $\sigma$-field $\mathcal{F}_\infty$. For this we need to use a limiting argument to consistently send $T \to \infty$ and obtain $(\mathsf{P}^h, L^h, B^h)$ as an appropriate limit of $(\mathsf{P}^h_T, L^h_T, B^h_T)$. Our proposal for this is as follows. For any process $h \in \mathcal{H}$ and an increasing sequence of finite times $\{T_n\}_{n\in\mathbb{N}}$, we repeatedly apply the Girsanov's theorem to obtain a family of consistent probability measures $\left\{\mathsf{P}^h_{T_n}, \mathcal{F}_{T_n} : n \in \mathbb{N}\right\}$, where $\mathsf{P}^h_{T_n} \sim \mathsf{P}$ on $\mathcal{F}_{T_n}$ for every $n \in \mathbb{N}$. In a similar vein we obtain the likelihood ratio process $\left\{L^h_{t,T_n} : t \leq T_n\right\}$ and the Brownian motion $\left\{B^h_{t,T_n} : t \leq T_n\right\}$ for every $n \in \mathbb{N}$. Next, we explain how to naturally define the limit of each three components.

(i) **Likelihood process limit:** Expression (3.3) implies that the sequence of likelihood processes are path-wise consistent with each other, i.e $L^h_{t,T_m} = L^h_{t,T_n}$ for every $t \leq$

[13] We refer to Karatzas and Shreve [2012] for the definition of progressive processes.

[14] The notion of integral depends on the context that could either be the path-wise Stieltjes integral or stochastic Itô integral.

$T_m \leq T_n$. Therefore, one can define the process $L^h$ on $[0, \infty)$ in a meaningful sense, such that its restriction to any finite interval coincides with the sequence of likelihood processes. This concludes the construction of the limit likelihood process. Importantly, this construction suggests that $L^h$ must be a martingale process with respect to $\mathsf{P}$ on $\mathbb{R}_+$. To see this, note that a bounded $h$ causes the *Novikov's* condition to hold, thereby $L^h_{T_n}$ would be an uniformly integrable martingale — on $[0, T_n]$ — with respect to $\mathsf{P}$ for every $n \in \mathbb{N}$. Because of the path-wise equivalence, this would immediately establish the martingale property of $L^h$ on $\mathbb{R}_+$.

(ii) **Probability measure limit:** First, recall that for every $n \in \mathbb{N}$, $\mathsf{P}^h_{T_n}$ is a probability measure on $\mathcal{F}_{T_n}$. Then, the path-wise consistency resulted from (3.3) implies that these measures indeed match each other, namely $\mathsf{P}^h_{T_n}(A) = \mathsf{P}^h_{T_m}(A)$ for every $A \in \mathcal{F}_{T_m}$ where $m \leq n$. Thus, we can apply theorem 4.2 in Parthasarathy [2005] that guarantees the existence of a *closing* probability measure $\mathsf{P}^h$ on $\mathcal{F}_\infty$ such that its restrictions to finite intervals coincide with the above sequence of probability measures, yet it need not be equivalent to $\mathsf{P}$ on $\mathcal{F}_\infty$. That is restricted to every finite $T$, $\mathsf{P}^h \sim \mathsf{P}$ on $\mathcal{F}_T$, but this may not be true on $\mathcal{F}_\infty$.

(iii) **Brownian motion limit:** Applying Girsanov's theorem lets us to deduce that $B^h_{T_n} := \{B^h_{t,T_n} : t \leq T_n\}$ is a Brownian motion under $\mathsf{P}^h_{T_n}$ on $[0, T_n]$ for every $n \in \mathbb{N}$. Since $\mathsf{P}^h \equiv \mathsf{P}^h_{T_n}$ on $[0, T_n]$, then it turns out that $B^h_{T_n}$ is also a Brownian motion under $\mathsf{P}^h$. Also note that the path-wise consistency holds for the sequence of Brownian motions, namely $B^h_{t,T_n} = B^h_{t,T_m}$ for all $t \leq T_m \leq T_n$. Therefore, in the same manner that we defined $L^h$ from $\{L^h_{T_n} : n \in \mathbb{N}\}$, we can define $B^h$ as the process on $\mathbb{R}_+$ such that its restrictions to any finite interval satisfy the properties of $\mathsf{P}^h$ Brownian motions.

The illustrated construction of $\left(\mathsf{P}^h, B^h\right)$ allows us to express the return process of the ambiguous project in term of $h$-Brownian motion:

$$\mathrm{d}y_{2,t} = \left[\mu_t\theta + \sigma\sqrt{\mu_t}h_t\right]\mathrm{d}t + \sigma\sqrt{\mu_t}\mathrm{d}B^h_t \tag{3.4}$$

The merit of weak formulation now becomes clear, where for every $\mu \in \mathcal{U}$ the return processes $\{y_1, y_2\}$ are essentially fixed, but the probability distribution that assigns weights to the subsets of sample paths is controlled by the choice of $h \in \mathcal{H}$. So in a sense the nature's move is to select the return's distribution not its sample paths.

Now that we know what is meant by $\mathsf{P}^h$ on $\mathcal{F}_\infty$ we can analyze both terms of (3.2) which are expectations under $\mathsf{P}^h$, and this will be the goal of next subsection.

## 3.3 Unravelling the payoff function

We begin the simplification of (3.2) by elaborating the second term, that is the entropy cost of ambiguity aversion. Recall that $\mathsf{P}$ and $\mathsf{P}^h$ need not necessarily be equivalent measures on $\mathcal{F}_\infty$, yet their restrictions $\{\mathsf{P}_t, \mathsf{P}_t^h\}$ are indeed equivalent probability measures on $\mathcal{F}_t$. Having that said, the relative discounted entropy is defined as

$$H\left(\mathsf{P}^h;\mathsf{P}\right) := \lim_{T\to\infty} \delta \int_0^T e^{-\delta t} H\left(\mathsf{P}_t^h;\mathsf{P}_t\right) \mathrm{d}t, \tag{3.5}$$

where $H\left(\mathsf{P}_t^h;\mathsf{P}_t\right) := \mathsf{E}^h\left[\log L_t^h\right]$. Expression (3.5), which is proposed in Hansen et al. [2006], presents a proxy for the discrepancy between two measures that are not necessarily equivalent on the terminal $\sigma$-field, and hence their relative entropy $H\left(\mathsf{P}_\infty^h;\mathsf{P}_\infty\right)$ could be infinite, but on each finite interval say $[0,t]$ they are equivalent $\mathsf{P}_t^h \sim \mathsf{P}_t$ and have finite relative entropy. Therefore, one shall hope that relation (3.5) is well-defined.

**Lemma 3.3.** *The discounted relative entropy in* (3.5) *is well-defined, namely for every* $h \in \mathcal{H}$ *it is finite and satisfies*

$$H\left(\mathsf{P}^h;\mathsf{P}\right) = \frac{1}{2}\mathsf{E}^h\left[\int_0^\infty e^{-\delta t} h_t^2 \mathrm{d}t\right] < \infty.$$

Roughly speaking, for the first component of the payoff function we need to take the expectation of $\mathrm{d}y_2$ under the measure $\mathsf{P}^h$. This is in our reach because we stated the dynamics of $y_2$ in terms of $B^h$ in (3.4). However, the drift term in $\mathrm{d}y_2$ contains the random variable $\theta$, that needs to be learned and projected onto the DM's information set. For this we present an optimal filtering result under each measure $\mathsf{P}^h$.

**Remark 3.4.** The DM's initial prior $p_0 = \mathsf{P}\left[\theta = \overline{\theta}\right]$ is unaffected under different probability distributions $\mathcal{P} = \left\{\mathsf{P}^h : h \in \mathcal{H}\right\}$. This is because the benchmark measure $\mathsf{P}$ and all its variations $\mathcal{P}$ agree on $\mathcal{F}_0$, resulted from $L_0^h = 1$ for every $h \in \mathcal{H}$.

In light of this remark, we want to continuously estimate and update the DM's posterior on $\theta$ based on her available information at every point in time. Her information set at time $t$ contains the path of output from each project $\left\{\left(y_{1,s}, y_{2,s}\right) : s \le t\right\}$, the history of her allocation process $\{\mu_s : s \le t\}$ and importantly the nature's moves up until time $t$, i.e $\{h_s : s \le t\}$. Note that at each time $t$, the DM's ambiguity is with regard to the future path of $h$, and she has no uncertainty about the history of nature's moves in the past. Some might not be willing to make this assumption about the ex-post observability of nature's moves to the DM. However, this is not an important assumption for two reasons. First, on the equilibrium path the DM knows the history of nature's past moves. Secondly, in

theory we can find the filtering equation under every possible history of nature's actions and then let the DM to pessimistically choose from this family of posteriors. In summary, the filtering problem that the DM faces at time $t$ is to update her posterior based on the available information set $\mathcal{F}_t^{y_1,y_2,\mu,h}$. Of secondary importance is to note that $y_1$ conveys no information about $\theta$, thus can be dropped out of information set.

**Definition 3.5.** For every $t \in (0,\infty)$, define $p_t^h := \mathsf{P}^h\left[\theta = \bar{\theta} \middle| \mathcal{F}_t^{y_2,\mu,h}\right]$ as the posterior probability and $m(p_t^h) = p_t^h\bar{\theta} + (1-p_t^h)\underline{\theta}$ as the conditional mean. At $t = 0$, let $p_t^h = p_0$ and $m(p_0^h) = m(p_0)$.

**Lemma 3.6** (Liptser and Shiryaev [2013] theorem 8.1)**.** *The conditional probability of the event $[\theta = \bar{\theta}]$ given the filtration $\mathbf{F}^{y_2,\mu,h}$ evolves according to the following stochastic differential equation:*

$$\mathrm{d}p_t^h = \frac{(\bar{\theta}-\underline{\theta})\sqrt{\mu_t}}{\sigma} p_t^h\left(1-p_t^h\right)\mathrm{d}\bar{B}_t^h \tag{3.6}$$

*Here $\left\{\bar{B}_t^h, \mathcal{F}_t^{y_2,\mu,h} : t \in \mathbb{R}_+\right\}$ is called the innovation process which is a Brownian motion under $\mathsf{P}^h$, and is characterized by $\mathrm{d}\bar{B}_t^h = \sigma^{-1}\sqrt{\mu_t}\left[\theta - m(p_t^h)\right]\mathrm{d}t + \mathrm{d}B_t^h$. As a result of this, the law of motion for $y_2$ would be*

$$\mathrm{d}y_{2,t} = \left[\mu_t m(p_t^h) + \sqrt{\mu_t}h_t\right]\mathrm{d}t + \sigma\sqrt{\mu_t}\mathrm{d}\bar{B}_t^h. \tag{3.7}$$

At this stage we have developed all the required tools to present the utility function in (3.2) in terms of initial belief and the players' actions. For this we define the infinite horizon payoff as the limit of finite horizon counterparts. The reason is that the constructed process $B^h$ is only Brownian motion over finite intervals, and we can not extend it to entire $\mathbb{R}_+$, unless we impose further restrictions on $\mathcal{H}$ and $\mathcal{U}$ to obtain the uniform integrability of likelihood processes, which we refrain to do. Therefore, inspired by (3.2) we define the utility of DM from taking action $\mu$ while nature chooses $h$ by

$$V(p;\mu,h) := \lim_{T\to\infty} \mathsf{E}^h\left[\delta\int_0^T e^{-\delta t}\left(\mathrm{d}y_{1,t} + \mathrm{d}y_{2,t} + \alpha H\left(P_t^h; P_t\right)\mathrm{d}t\right)\right]. \tag{3.8}$$

**Proposition 3.7.** *For every choice of $\mu \in \mathcal{U}$ and $h \in \mathcal{H}$, the net discounted average payoff defined in (3.8) can be expressed as:*

$$V(p;\mu,h) = \mathsf{E}^h\left[\delta\int_0^\infty e^{-\delta t}\left((1-\mu_t)r + \mu_t m(p_t^h) + \sigma\sqrt{\mu_t}h_t + \frac{\alpha}{2\delta}h_t^2\right)\mathrm{d}t\right] \tag{3.9}$$

This proposition serves us well, because the integrand is now $\mathbf{F}^{y_2,\mu,h}$-progressively measurable, that in turn allows us to perform a dynamic programming scheme to express the

*value function* in terms of the current belief, and this will be the goal of next section.

# 4 Dynamic programming analysis

Our analysis so far offers expression (3.9) as the DM's payoff in the two-player differential game against the nature. For any point of time, say $t \in \mathbb{R}_+$, define the expected continuation value conditioned on $\mathcal{G}_t := \mathcal{F}_t^{y_2,\mu,h}$ as

$$J(p,t;\mu,h) := \mathsf{E}^h\left[\delta\int_t^\infty e^{-\delta s}\left((1-\mu_s)r+\mu_s m(p_s^h)+\sigma\sqrt{\mu_s}h_s+\frac{\alpha}{2\delta}h_s^2\right)\mathrm{d}s\,\middle|\,\mathcal{G}_t\right]. \tag{4.1}$$

In that $p$ is the time $t$ value of the state process $p_t^h$. For every $h \in \mathcal{H}$ the process $\bar{B}^h$ as well as $p^h$ are time *homogeneous* Markov diffusions. Furthermore, the players' action spaces at the time $t$ sub-game — $\mathcal{U}_t$ and $\mathcal{H}_t$ resp. for the DM and the nature — are essentially isomorphic to $\mathcal{U}$ and $\mathcal{H}$. These two premises imply that max-min value of the game for the DM, i.e $\sup_{\mu\in\mathcal{U}_t}\inf_{h\in\mathcal{H}_t} J(p,t;\mu,h)$, is time homogeneous. Specifically, there exists a value function $v(p)$ such that

$$\sup_{\mu\in\mathcal{U}_t}\inf_{h\in\mathcal{H}_t} J(p,t;\mu,h) = e^{-\delta t}v(p) \tag{4.2}$$

Our goal in the next theorem is to present a *verification* result for the value function. In particular, next theorem states that if $v(\cdot)$ is a smooth enough function satisfying a certain HJBI equation, then it is indeed the value function in (4.2).

**Theorem 4.1.** *Suppose $v \in C_b^2(0,1)$[15] is the unique solution to the following HJBI equation:*

$$v(p) = \sup_{\mu\in[0,1]}\inf_{h\in\mathbb{R}}\left\{(1-\mu)r+\mu m(p)+\sigma\sqrt{\mu}h+\frac{\alpha}{2\delta}h^2+\frac{\mu}{2\delta}\Phi(p)v''(p)\right\}, \tag{4.3}$$

*where $\Phi(p) := \sigma^{-2}(\bar{\theta}-\underline{\theta})^2p^2(1-p)^2$. Then, $v$ is indeed the value function in (4.2). In the equilibrium, the worst-case density generator is $h^* = -\alpha^{-1}\sigma\delta\sqrt{\mu^*}$, where $\mu^*$ is the DM's best response in*

$$v(p) = \sup_{\mu\in[0,1]}\left\{(1-\mu)r+\mu m(p)-\frac{\sigma^2\delta}{2\alpha}\mu+\frac{\mu}{2\delta}\Phi(p)v''(p)\right\}. \tag{4.4}$$

This theorem only sets out a partial characterization for the value function, in the sense that it doesn't claim the value function is a *generalized* solution to the HJBI equation. However, we believe that with some nontrivial work one can adapt the seminal work of Fleming and Souganidis [1989] to the weak-formulation and prove that the value function

[15] Space of bounded functions which are twice continuously differentiable on $(0,1)$.

is the unique viscosity solution of (4.3), henceforth we assume even if $v \notin C_b^2(0,1)$ it still satisfies (4.4) in the viscosity sense.

As stated in previous theorem, on the equilibrium path of the game, DM knows the best response of the nature, that is $h(\mu) = -\alpha^{-1}\sigma\delta\sqrt{\mu}$. Therefore, her posterior process follows that of (3.6) for the prescribed $h(\mu)$. Importantly, this means at the equilibrium the DM is no longer concerned about all possible distributions of past shocks. The one that has been picked by the nature is known to the DM on the equilibrium path, which gives rise to the unique law of motion for the posterior belief. Note that, this does not mean that ambiguity is mitigated on the equilibrium path. However, it simply means that similar to the static decision making, where the ambiguity averse agent first perceives the worst case distribution from her multiple prior set, and then responds back, here also she forms her belief and react based on the worst case distribution choice by the nature. Henceforth, by $p$ in (4.4) and in the rest of the paper we mean the equilibrium posterior value, or often for brevity is simply referred as *belief*.

Note that the *rhs* of (4.4) is linear in $\mu$. This is in part due to the effect of $\sqrt{\mu}$ as the volatility term in the ambiguous arm. Consequently, the DM's optimal strategy at every point in time is to either *explore* the ambiguous arm or *exploit* the safe arm[16]. As a result, the DM's value function satisfies the following variational relation:

$$v(p) = \max\left\{r, m(p) - \frac{\sigma^2\delta}{2\alpha} + \frac{1}{2\delta}\Phi(p)v''(p)\right\} \tag{4.5}$$

In the economic terms, $r$ is the DM's reservation value, which can always be achieved regardless of her experimentation strategy. The term $m(p)$ is the expected rate of return from pulling the second arm when the current belief on $\theta$ is $p$. The important term in expression (4.5) is $\sigma^2\delta/2\alpha$, which we call it *ambiguity cost*. Higher ambiguity aversion, translated to lower $\alpha$, implies higher incurred cost upon pulling the ambiguous arm. Lastly, $\frac{1}{2\delta}\Phi(p)v''(p)$ is the continuation payoff that the DM could expect by holding on to the second arm. We postpone a more elaborate set of analytical results on the value function to the next subsection and instead present the intuition behind the DM's optimal strategy.

**Lemma 4.2.** *The DM's optimal allocation choice with ambiguity aversion $\alpha$ admits the*

[16]The trade-off between exploration vs. exploitation has studied in different context. For one we can point to Manso [2011] that explains such a trade-off for the financial incentives in entrepreneurship.

*following representation:*

$$\mu^*(p) = \begin{cases} 1 & \text{if } \frac{1}{2\delta}\Phi(p)v''(p) - \frac{\sigma^2\delta}{2\alpha} > r - m(p) \\ \in [0,1] & \text{if } \frac{1}{2\delta}\Phi(p)v''(p) - \frac{\sigma^2\delta}{2\alpha} = r - m(p) \\ 0 & \text{otherwise} \end{cases} \tag{4.6}$$

This result is the analogue of lemma 4 in Bolton and Harris [1999] tailored to capture the ambiguity aversion. One shall think of $r - m(p)$ as the opportunity cost of experimentation that the DM incurs by not choosing the safe arm. Therefore, she only selects the second project when the continuation value of experimentation adjusted by the ambiguity price exceeds its opportunity cost. Particularly, whenever the two values match, the DM can pursue a *mixed strategy*, in that she can allocate her resources between two arms in any arbitrary proportions. However, the Lebesgue measure of the time duration on which she chooses the mixed strategy is zero, precisely because $p$ follows a diffusion process and the middle case in (4.6) never happens $\mathsf{dP} \times \mathsf{Leb}$-a.e. The ambiguity aversion essentially creates a situation in that the DM thinks that upon the continuation she will have to face with the most destructive types of shock distribution, and this already lowers the value of experimentation. Importantly, this loss is independent of the current belief level, and shall be viewed as a fixed cost that ambiguity averse agent must be compensated for to undertake the second project.

# 5 Properties of the value function and comparative statics

In this section we propose closed-form expression for the value function and present sharp comparative statics with respect to ambiguity aversion index $\alpha$.

**Theorem 5.1.** *On the equilibrium path the DM's follows a cut-off experimentation strategy. In particular, there exists $\bar{p} \in [0,1]$ such she selects the safe arm if and only if her posterior belief drops below $\bar{p}$. Further, the value function $v$ is convex on $[0,1]$.*

A substantive result of convexity is that even in the presence of ambiguity aversion the marginal value of *good news* about the second project is increasing.

Next, we want to find a closed-form expression for the value function and particularly the cut-off probability $\bar{p}$. For this we make a technical assumption that turns out to be necessary and sufficient for existence of $\bar{p}$ in $(0,1)$. Namely, we exclude the case $\bar{p} = 0$ where DM always pulls the second arm, and $\bar{p} = 1$ where she never does.

**Assumption 5.2.** Define $\eta := \frac{r-\underline{\theta}}{\bar{\theta}-\underline{\theta}} + \frac{\sigma^2\delta}{2\alpha(\bar{\theta}-\underline{\theta})}$. Then we assume $\eta < 1$.

As becomes clear later, one can think of $\eta$ as a lower bound on $\bar{p}$. Therefore $\eta > 1$ essentially means that DM never selects the ambiguous arm. This is due to a combination

of two forces, namely a large ratio of safe to ambiguous return — that is the first term in $\eta$ — and high normalized ambiguity cost — that is the second term in $\eta$ — which prevents the DM from exploring the second arm. Assumption 5.2 not only ensures that $\bar{p} < 1$, but as it will turn out it implies $\bar{p} > 0$. Having made this assumption, on *exploration region* $(\bar{p}, 1]$ the following differential equation holds:

$$v(p) = m(p) - \frac{\sigma^2\delta}{2\alpha} + \frac{1}{2\delta}\Phi(p)v''(p) \tag{5.1}$$

That has a general solution form[17]

$$v(p) = m(p) - \frac{\sigma^2\delta}{2\alpha} + cp^{1-\lambda}(1-p)^{\lambda}, \quad \text{on } p \in (\bar{p}, 1]. \tag{5.2}$$

Here $c$ is a constant determined from the boundary condition and $\lambda = \frac{1+\sqrt{1+4\delta\varphi^{-2}}}{2}$, where $\varphi := (\bar{\theta} - \underline{\theta})/\sigma\sqrt{2}$. The *value-matching* (or equivalently *no-arbitrage*) condition implies that the DM should be indifferent between choosing any of the two arms at $p = \bar{p}$. Therefore, $v(\bar{p}) = r$ that yields to

$$v(p) = m(p) - \frac{\sigma^2\delta}{2\alpha} + \left(r - m(\bar{p}) + \frac{\sigma^2\delta}{2\alpha}\right)\frac{p^{1-\lambda}(1-p)^{\lambda}}{\bar{p}^{1-\lambda}(1-\bar{p})^{\lambda}}, \quad \forall p \in [\bar{p}, 1]. \tag{5.3}$$

The DM faces a *free-boundary* problem, namely she needs to find the optimal cut-off $\bar{p}$. For that we need to apply the *smooth-pasting*[18] condition that imposes the continuity of directional derivatives at $\bar{p}$, i.e $v'(\bar{p}^-) = v'(\bar{p}^+)$. Assumption 5.2 with some amount of algebra yields to the following expression for the cut-off probability:

$$\bar{p} = \frac{(\lambda - 1)\eta}{\lambda - \eta} \tag{5.4}$$

It is positive because $\eta < 1 \leq \lambda$, and is less than one again because $\eta < 1$. This observation now supports making assumption 5.2.

**Some comparative statics.** The cut-off value is lower-bounded by $\eta$. Further, it is increasing in $\eta$. Expression (5.4) provides us with a sharp characterization of the cut-off value, and one could perform a number of comparative statics on $\bar{p}$ with respect to the parameters of the model. Here, we only point to two interesting ones. First, and more important is the effect of ambiguity on cut-off value. As DM becomes more ambiguity averse, namely as $\alpha$ becomes smaller, the value of $\bar{p}$ increases unambiguously. This confirms our intuition that

[17] Polyanin and Zaitsev [2017] page 547.
[18] Dixit [2013].

a more ambiguity averse DM is more conservative and explores less. Expression (5.4) offers a fine indicator on the *extent* of this under-exploration. The second channel is the effect of $\overline{\theta} - \underline{\theta}$, that represents the *range* of possible return rates under the second arm. As this range shrinks to zero, the ambiguity cost is amplified more intensely, and DM will have less incentive to pick the second project.

As a last note in this section we point out to a concern on the entangled effects of $\sigma$ and $\alpha$. One might wonder that what we refer as the ambiguity aversion parameter, i.e $\alpha^{-1}$, can be dissolved in volatility $\sigma$, and thus can never be identified separately even with infinite amount of data. However, this is not true, as we can offer an identification scheme that disentangles $\alpha$ from $\sigma$. Suppose that all other parameters are identified, namely $r, \delta$ and $\{\bar{\theta}, \underline{\theta}\}$. Then, a continuous stream of agent's belief process would let us to compute the quadratic variation $\langle p, p\rangle = (\bar{\theta} - \underline{\theta})^2 p^2 (1-p)^2/\sigma^2$ from (3.6). Further, by spotting the point where she stops the exploration and pulls the safe arm we can back out $\bar{p}$. These two equations can lead us to uniquely identify $\sigma$ and $\alpha$.

# 6 Value of unambiguous information

In this section we aim to study the value of information with respect to which the DM holds no ambiguity. Practically, one can think of a scenario in which the experimentation unit hires an expert to continuously provide her opinion about the *true* rate of return of the ambiguous arm. Some questions naturally arise in this context. For example what is the fair price of such service? Or, how much must the expert be compensated for providing such information? When should the experimentation unit who faces ambiguity hire this expert?

To answer such questions, let $x_t$ be the information that the expert releases at time $t$ about $\theta$, which in its simplest case can be thought as the noisy signal of $\theta$, namely:

$$\mathrm{d}x_t = \theta \mathrm{d}t + \gamma \mathrm{d}W_t \tag{6.1}$$

In this expression $W$ is a $\mathbf{F}$-Brownian motion under the benchmark measure $\mathsf{P}$ and is independent of $B$ and $\theta$. Further, $\gamma$ is the constant volatility that represents the level of DM's confidence in the expert's information. Therefore, the DM can use this signal in addition to the second arm's payoff process to update her belief about $\theta$. Obviously, this new source of information improves the precision of the filtering process, in the sense that it lowers the conditional variance of estimated $\theta$ at every point in time. The law of motion for the new posterior process with the presence of unambiguous information source follows the logic of lemma 3.6:

$$\mathrm{d}p_t^h = p_t^h(1-p_t^h)\left(\bar{\theta} - \underline{\theta}\right)\left[\frac{\sqrt{\mu_t}}{\sigma}\mathrm{d}\bar{B}_t^h + \frac{1}{\gamma}\mathrm{d}\overline{W}_t\right] \tag{6.2}$$

Here $\bar{B}$ and $\overline{W}$ are independent $\mathbf{F}^{y_2,\mu,h,x}$-Brownian motions under $\mathsf{P}^h$. Now we can state the counterpart of theorem 4.1 in this case, where its proof is almost the same as theorem 4.1.

**Proposition 6.1.** *Suppose $\tilde{v} \in C_b^2(0,1)$ is the unique solution to the following HJBI equation:*

$$\tilde{v}(p) = \sup_{\mu\in[0,1]} \inf_{h\in\mathbb{R}} \left\{ (1-\mu)r + \mu m(p) + \sqrt{\mu}\sigma h + \frac{\alpha}{2\delta}h^2 + \frac{1}{2\delta}\left(\mu\Phi(p;\sigma) + \Phi(p;\gamma)\right)\tilde{v}''(p) \right\} \tag{6.3}$$

*In that $\Phi(p;s) := \frac{(\bar{\theta}-\underline{\theta})^2}{s^2}p^2(1-p)^2$. Then, $\tilde{v}$ is indeed the value function in presence of unambiguous information $x$. In the equilibrium, the worst-case density generator is $h^* = -\alpha^{-1}\sigma\delta\sqrt{\mu^*}$, where $\mu^*$ is the DM's best response solving:*

$$\tilde{v}(p) = \sup_{\mu\in[0,1]} \left\{ (1-\mu)r + \mu m(p) - \frac{\sigma^2\delta}{2\alpha}\mu + \frac{1}{2\delta}\left(\mu\Phi(p;\sigma) + \Phi(p;\gamma)\right)\tilde{v}''(p) \right\} \tag{6.4}$$

Similar to the case with no source of unambiguous information, one can show that the value function is non-decreasing in $p$ and there is a cut-off rule for the optimal experimentation strategy. Let us denote the new cut-off in the presence of unambiguous information with $\tilde{p}$. Then, the value function satisfies the following relation:

$$\tilde{v}(p) = \begin{cases} r + \delta^{-1}\varphi(\gamma)^2 p^2(1-p)^2\tilde{v}''(p) & p < \tilde{p} \\ m(p) - \frac{\sigma^2\delta}{2\alpha} + \delta^{-1}\left(\varphi(\sigma)^2 + \varphi(\gamma)^2\right)\tilde{v}''(p) & p > \tilde{p} \end{cases} \tag{6.5}$$

In that we define $\varphi(s) = \left(\bar{\theta} - \underline{\theta}\right)/s\sqrt{2}$, where $s \in \{\sigma, \gamma\}$. The top term in (6.5) relates to the region where DM selects the safe arm. Importantly, on this region her payoff is no longer $r$, but has a continuation component that arises from the free information $x$. In the case without this source, once the DM switches to the safe arm, she will never have the chance to acquire information about $\theta$, thereby her payoff will stuck at $r$ forever. However, in the current situation, the news about $\theta$ can still be flowing without DM pulling the second arm, and in the case of *good* news, she would expect to switch back to the second arm. This effect creates the continuation incentives for the DM on the region $(0, \tilde{p})$. At $\tilde{p}$ the continuity condition must hold so any of the two regions in (6.5) could be enclosed. The solution to this piece-wise ordinary differential equation is

$$\tilde{v}(p) = \begin{cases} r + c_1 p^{\lambda_1}(1-p)^{1-\lambda_1} & p < \tilde{p} \\ m(p) - \frac{\sigma^2\delta}{2\alpha} + c_2 p^{1-\lambda_2}(1-p)^{\lambda_2} & p > \tilde{p}, \end{cases} \tag{6.6}$$

where $\lambda_1 = \frac{1+\sqrt{1+4\delta\varphi(\gamma)^{-2}}}{2}$ and $\lambda_2 = \frac{1+\sqrt{1+4\delta\left(\varphi(\sigma)^2+\varphi(\gamma)^2\right)^{-1}}}{2}$. There are essentially three parameters to be determined, i.e $(c_1, c_2, \tilde{p})$. The optimal choice of DM is to select these constants

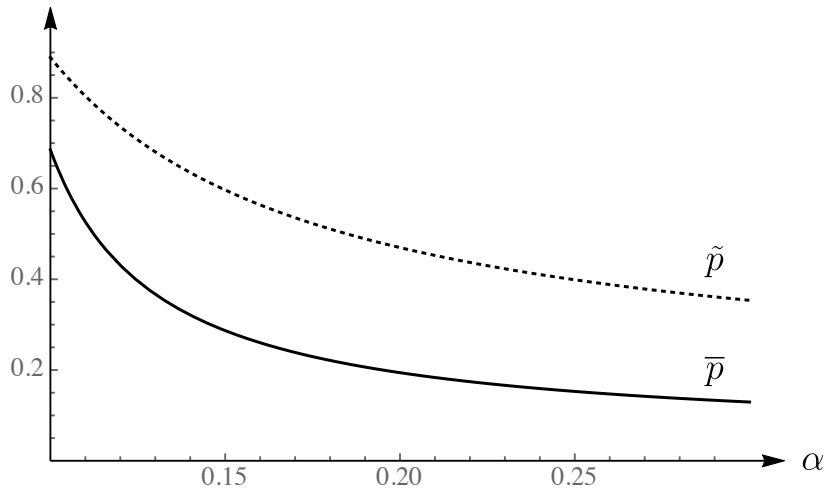


Figure 1: Cut-off values

$$\left[r = 0.2, \underline{\theta} = 0, \overline{\theta} = 1, \delta = 0.9, \sigma = 0.4, \gamma = 0.3\right]$$

so that the three conditions, namely *value-matching* (continuity), *smooth-pasting* (continuity of first derivative) and *super-contact* (continuity of second derivative) hold together. The derivations for this are presented in A.5. It turns out the new cut-off probability under unambiguous information source is

$$\tilde{p} = \frac{(\Lambda - 1)\eta}{\Lambda - \eta}, \quad \text{for } \Lambda := 1 + \lambda_1 \frac{\sigma^2}{\gamma^2} + (\lambda_2 - 1)\left(1 + \frac{\sigma^2}{\gamma^2}\right). \tag{6.7}$$

**Proposition 6.2.** *The experimentation cut-off rises unambiguously when there is an unambiguous information source, namely $\tilde{p} \geq \bar{p}$ for all combinations of the variables in the model.*

The content behind this proposition is that the unambiguous source of information in effect raises the bar for exploration, that in turn means DM demands more confidence for selecting the second project. This is very much due to the free information that DM can acquire about $\theta$ without pulling the ambiguous arm. In the standard case, the only way to learn about the quality of the second project is to spend some time exploring that. Therefore, the DM is more willing to sacrifice the certain payoff of the first project to learn about the second one, whereas in the current case she can wait longer for the good news (and exploit the first arm meanwhile) to choose the second arm. In this spirit, as depicted in figure 2 the cut-off value rises unambiguously due to the provision of the new information source (i.e $\tilde{p} > \bar{p}$). Also it shows that in both environments the exploration threshold falls as the DM becomes less ambiguity averse, meaning larger values of $\alpha$.

One can think of a situation where the provider of this new source of information is

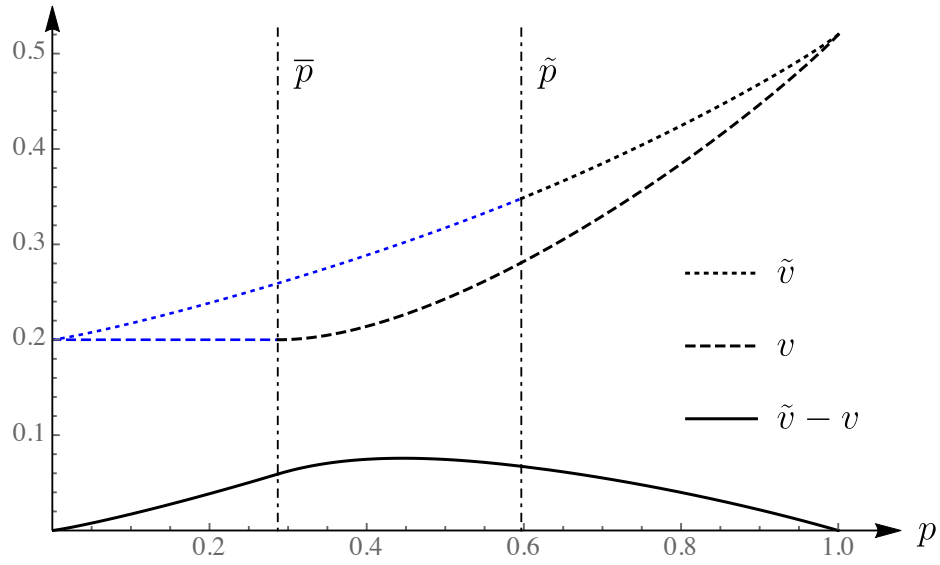


Figure 2: Created surplus and value functions

$$\left[r = 0.2, \underline{\theta} = 0, \overline{\theta} = 1, \delta = 0.9, \sigma = 0.4, \gamma = 0.3, \alpha = 0.14\right]$$

strategic and can charge the DM for the service. Then naturally the maximum price that she can charge is $\tilde{v}(p) - v(p)$, which corresponds to extracting all the surplus from the DM. From the social welfare standpoint the $p$ that maximizes the surplus shall be treated as a benchmark for decision to hire the expert. We refer to $\tilde{v}(p) - v(p)$ as the created surplus due the expert opinion. It is obviously positive and continuous everywhere, and is increasing over $[0, \bar{p}]$. Also as $p \to 1$ it decays to zero faster than $(1-p)^{\lambda_2}$. Therefore, the maximum created surplus occurs at a *moderate* belief value $p^*$, where $p^* > \bar{p}$ but is not also very close to one. Figure 1 presents both value functions, and the created surplus. In that the blue segment of each curve points to the region where the DM pulls the safe arm. We end this section with a remark about the most efficient time to hire an expert.

**Remark 6.3.** The above analysis implies that it is most beneficial for the experimentation unit to hire an expert when otherwise they would select the ambiguous arm in spite of strong enough evidence and belief.

# 7 Concluding remarks

How does a decision maker who is *uncertain* about the payoff distribution of two alternative choices operate the dynamics of experimentation? Understanding how an ambiguity averse agent values a project and determining the *price* of ambiguity are particularly important when the experimentation task is delegated to such agent. In this paper, we develop a dy-

namic decision making framework that offers closed-form characterizations for the agent's optimal strategy as well as her valuation. Specifically, we assumed the DM has Multiplier preferences, that consists of two components. The discounted expected future return from both arms, and a penalty term that captures the extent of perturbation of probability specification relative to the benchmark model. We framed the decision making environment as a two-player differential game that DM plays against the nature, and found a closed-form expression for DM's value function in terms of her belief. Also, we have shown that in the equilibrium her optimal strategy is to select the safe arm of the project whenever her belief drops below a certain threshold, the value of which is controlled by all the parameters of the model and specifically the ambiguity aversion index. Our analysis offers sharp results on how much an ambiguity averse DM must be compensated to act as if she is not subject to ambiguity. In particular, one can send $\alpha \to \infty$ in the results of section 5 to predict the behavior of an ambiguity neutral agent. Finally, we explored the effect of an unambiguous constantly flowing information source in the dynamics of experimentation. It turned out that the exploration cut-off rises as a result of such provision, namely the DM waits longer to receive good news about the ambiguous arm of the project. We investigated the generated surplus due to this additional source and offered policy analysis on the efficient time to recruit an external expert to guide the experimentation process.

## References


Christopher M Anderson. Ambiguity aversion in multi-armed bandit problems. *Theory and decision*, 72(1):15–33, 2012.

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

Alessandro Bonatti and Johannes Hörner. Learning to disagree in a game of experimentation. *Journal of Economic Theory*, 169:234–269, 2017.

Felipe Caro and Aparupa Das Gupta. Robust control of the multi-armed bandit problem. *Annals of Operations Research*, pages 1–20, 2013.

Xue Cheng and Frank Riedel. Optimal stopping under ambiguity in continuous time. *Mathematics and Financial Economics*, 7(1):29–68, 2013.

Avinash Dixit. *The art of smooth pasting*. Routledge, 2013.

Larry G Epstein and Shaolin Ji. Optimal learning and ellsberg's urns. *arXiv preprint arXiv:1708.01890*, 2017.

Larry G Epstein and Martin Schneider. Recursive multiple-priors. *Journal of Economic Theory*, 113(1):1–31, 2003.

Larry G Epstein and Martin Schneider. Learning under ambiguity. *The Review of Economic Studies*, 74(4):1275–1303, 2007.

Wendell H Fleming and Panagiotis E Souganidis. On the existence of value functions of two-player, zero-sum stochastic differential games. *Indiana University Mathematics Journal*, 38(2):293–314, 1989.

Itzhak Gilboa and David Schmeidler. Maxmin expected utility with non-unique prior. *Journal of mathematical economics*, 18(2):141–153, 1989.

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

Lars Peter Hansen and Thomas J Sargent. Robust control and model uncertainty. *The American Economic Review*, 91(2):60–66, 2001.

Lars Peter Hansen and Thomas J Sargent. Robustness and ambiguity in continuous time. *Journal of Economic Theory*, 146(3):1195–1223, 2011.

Lars Peter Hansen, Thomas J Sargent, Gauhar Turmuhambetova, and Noah Williams. Robust control and model misspecification. *Journal of Economic Theory*, 128(1):45–90, 2006.

Paul Heidhues, Sven Rady, and Philipp Strack. Strategic experimentation with private payoffs. *Journal of Economic Theory*, 159:531–551, 2015.

Ioannis Karatzas and Steven Shreve. *Brownian motion and stochastic calculus*, volume 113. Springer Science & Business Media, 2012.

Godfrey Keller and Sven Rady. Optimal experimentation in a changing environment. *The review of economic studies*, 66(3):475–507, 1999.

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

Michael Jong Kim and Andrew EB Lim. Robust multiarmed bandit problems. *Management Science*, 62(1):264–285, 2015.

Jian Li. The k-armed bandit problem with multiple priors. *Journal of Mathematical Economics*, 80:22–38, 2019.

Robert S Liptser and Albert N Shiryaev. *Statistics of random Processes: I. general Theory*, volume 5. Springer Science & Business Media, 2013.

Fabio Maccheroni, Massimo Marinacci, and Aldo Rustichini. Ambiguity aversion, robustness, and the variational representation of preferences. *Econometrica*, 74(6):1447–1498, 2006a.

Fabio Maccheroni, Massimo Marinacci, and Aldo Rustichini. Dynamic variational preferences. *Journal of Economic Theory*, 128(1):4–44, 2006b.

Gustavo Manso. Motivating innovation. *The Journal of Finance*, 66(5):1823–1860, 2011.

Massimo Marinacci. Learning from ambiguous urns. *Statistical Papers*, 43(1):143–151, 2002.

Robert J Meyer and Yong Shi. Sequential choice under ambiguity: Intuitive solutions to the armed-bandit problem. *Management Science*, 41(5):817–834, 1995.

Jianjun Miao and Alejandro Rivera. Robust contracts in continuous time. *Econometrica*, 84 (4):1405–1440, 2016.

Kalyanapuram Rangachari Parthasarathy. *Probability measures on metric spaces*, volume 352. American Mathematical Soc., 2005.

Andrei D Polyanin and Valentin F Zaitsev. *Handbook of Ordinary Differential Equations: Exact Solutions, Methods, and Problems*. Chapman and Hall/CRC, 2017.

Frank Riedel. Optimal stopping with multiple priors. *Econometrica*, 77(3):857–908, 2009.

Paul Viefers. Should i stay or should i go?-a laboratory analysis of investment opportunities under ambiguity. *Working Paper*, 2012.

Martin L Weitzman. Optimal search for the best alternative. *Econometrica: Journal of the Econometric Society*, pages 641–654, 1979.

# A Appendix

## A.1 Proof of lemma 3.3

For every finite $T$ and $h \in \mathcal{H}$ the integral can be simplified as:

$$\begin{aligned}\delta \int_0^T e^{-\delta t} H\left(\mathsf{P}_t^h; \mathsf{P}_t\right) \mathrm{d}t &= \delta \int_0^T e^{-\delta t} \mathsf{E}^h \left[\log L_t^h\right] \mathrm{d}t \\ &= \delta \int_0^T e^{-\delta t} \mathsf{E}^h \left[(h \cdot B)_t - \frac{1}{2}\left(h^2 \cdot \imath\right)_t\right] \mathrm{d}t \\ &= \delta \int_0^T e^{-\delta t} \mathsf{E}^h \left[\left(h \cdot B^h\right)_t + \frac{1}{2}\left(h^2 \cdot \imath\right)_t\right] \mathrm{d}t\end{aligned} \tag{A.1}$$

Since $\{B_t^h : t \leq T\}$ is $\mathsf{P}_T^h$-Brownian motion and $h$ is bounded, the first term in above has zero expectation, leaving us only with the second term, for which integration by part yields

$$\mathsf{E}^h \left[\frac{\delta}{2} \int_0^T e^{-\delta t} \left(h^2 \cdot \imath\right)_t \mathrm{d}t\right] = \mathsf{E}^h \left[-\frac{1}{2} e^{-\delta T} (h^2 \cdot \imath)_T + \frac{1}{2} \int_0^T e^{-\delta t} h_t^2 \mathrm{d}t\right].$$

The first term inside expectation is uniformly bounded over $\Omega \times \mathbb{R}_+$ and goes to zero in a point-wise sense as $T \to \infty$. Therefore,

$$H\left(\mathsf{P}^h; \mathsf{P}\right) = \frac{1}{2} \lim_{T \to \infty} \mathsf{E}^h \left[\int_0^T e^{-\delta t} h_t^2 \mathrm{d}t\right] = \frac{1}{2} \mathsf{E}^h \left[\int_0^\infty e^{-\delta t} h_t^2 \mathrm{d}t\right],$$

where in the last relation we used the monotone convergence theorem. The limit is finite due to the boundedness of $h \in \mathcal{H}$. It is worthwhile to point out that for integrals with finite upper limit $T$ we appeal to $\mathsf{P}_T^h$, and for the infinite time integral we use $\mathsf{P}^h$. This replacement does not cause any problem because of the consistency of $\{\mathsf{P}_T^h : T \in \mathbb{R}_+\}$ with $\mathsf{P}^h$ as explained in item (ii) of subsection 3.2.

## A.2 Proof of proposition 3.7

Let us define

$$V^T(p; \mu, h) := \mathsf{E}^h \left[\delta \int_0^T e^{-\delta t} \left(\mathrm{d}y_{1,t} + \mathrm{d}y_{2,t} + \alpha H\left(P_t^h; P_t\right) \mathrm{d}t\right)\right]. \tag{A.2}$$

For the first two components of (A.2), one just need to recall that over every finite interval $[0, T]$, the pair $\left\{B_t^h, \mathcal{F}_t : t \leq T\right\}$ is a Brownian motion under $\mathsf{P}^h$. Consequently, stochastic integrals of bounded processes with respect to that are martingales and hence average out

to zero. Using equation (3.7) yields to

$$\begin{aligned}\mathsf{E}^h\left[\delta\int_0^T e^{-\delta t}\left(\mathrm{d}y_{1,t}+\mathrm{d}y_{2,t}\right)\right] &= \mathsf{E}^h\left[\delta\int_0^T e^{-\delta t}\left\{\left((1-\mu_t)r+\mu_t m(p_t^h)+\sigma\sqrt{\mu_t}h_t\right)\mathrm{d}t+\sigma\sqrt{\mu_t}\,d\bar{B}_t^h\right\}\right]\\ &= \mathsf{E}^h\left[\delta\int_0^T e^{-\delta t}\left((1-\mu_t)r+\mu_t m(p_t^h)+\sigma\sqrt{\mu_t}h_t\right)\mathrm{d}t\right]\end{aligned} \tag{A.3}$$

The entropy component of the integrand in (A.2) has already been analyzed in the proof of lemma 3.3 and specifically in equation (A.1). Combining that analysis with (A.3) leads to

$$\begin{aligned}V^T(p;\mu,h) = \mathsf{E}^h&\left[\delta\int_0^T e^{-\delta t}\left((1-\mu_t)r+\mu_t m(p_t^h)+\sigma\sqrt{\mu_t}h_t+\frac{\alpha}{2\delta}h_t^2\right)\mathrm{d}t\right]\\ &-\frac{1}{2}e^{-\delta T}\mathsf{E}^h\left[(h^2\cdot\imath)_T\right]\end{aligned} \tag{A.4}$$

Since $h\in\mathcal{H}$ is bounded, the second term in (A.4) vanishes as $T\to\infty$. For the first term in (A.4) we apply the dominated convergence theorem and use the fact that $\left\{\mathsf{P}_T^h : T\in\mathbb{R}_+\right\}$ is consistent with $\mathsf{P}^h\in\Delta\left(\Omega,\mathcal{F}_\infty\right)$[19] — as explained in (ii) of subsection 3.2 — while sending $T\to\infty$. This concludes the proof of $\lim_{T\to\infty}V^T(p;\mu,h)=V(p;\mu,h)$, thereby leading to (3.9).

### A.3 Proof of theorem 4.1

The presumption is that $v\in C_b^2(0,1)$ solves equation (4.3) and the goal is to prove that $v$ is indeed the correct value function in (4.2). For this let's define $g(\mu,h,p):=(1-\mu)r+\mu m(p)+\sigma\sqrt{\mu}h+\frac{\alpha}{2\delta}h^2$, and set

$$G_t(\mu,h):=\delta\int_0^t e^{-\delta s}g(\mu_s,h_s,p_s^h)\mathrm{d}s+e^{-\delta t}v(p_t^h).$$

Recall that $p_t^h$ follows the diffusion process $\mathrm{d}p_t^h=\sqrt{\Phi(p_t^h)\mu_t}\mathrm{d}\bar{B}_t^h$, where $\bar{B}^h$ is $\mathcal{G}$-Brownian motion under $\mathsf{P}^h$ over any finite horizon. Note that $\bar{B}^h$ is locally square integrable thus we can apply the Ito's lemma:

$$\begin{aligned}\mathrm{d}G_t &= \delta e^{-\delta t}g(\mu_t,h_t,p_t^h)\mathrm{d}t-\delta e^{-\delta t}v(p_t^h)\mathrm{d}t+e^{-\delta t}v'(p_t^h)\mathrm{d}p_t^h+e^{-\delta t}\frac{1}{2}v''(p_t^h)\mathrm{d}\langle p^h,p^h\rangle_t\\ &= \delta e^{-\delta t}\left[g(\mu_t,h_t,p_t^h)-v(p_t^h)+\frac{\mu_t}{2\delta}\Phi(p_t^h)v''(p_t^h)\right]\mathrm{d}t+e^{-\delta t}v'(p_t^h)\sqrt{\Phi(p_t^h)\mu_t}\mathrm{d}\bar{B}_t^h\end{aligned} \tag{A.5}$$

[19] $\Delta\left(\Omega,\mathcal{F}\right)$ denotes the set of all probability measures on the measure space $(\Omega,\mathcal{F})$.

For $\mu = \mu^*$ as prescribed in the theorem and arbitrary $h \in \mathcal{H}$ we have

$$\begin{aligned}
\mathrm{d}G_t(\mu^*, h) &= \delta e^{-\delta t}\left[g(\mu_t^*, h_t, p_t^h) - v(p_t^h) + \frac{\mu_t^*}{2\delta}\Phi(p_t^h)v''(p_t^h)\right]\mathrm{d}t + e^{-\delta t}v'(p_t^h)\sqrt{\Phi(p_t^h)\mu_t^*}\mathrm{d}\bar{B}_t^h \\
&\geq \delta e^{-\delta t}\left[g(\mu_t^*, h_t^*, p_t^h) - v(p_t^h) + \frac{\mu_t^*}{2\delta}\Phi(p_t^h)v''(p_t^h)\right]\mathrm{d}t + e^{-\delta t}v'(p_t^h)\sqrt{\Phi(p_t^h)\mu_t^*}\mathrm{d}\bar{B}_t^h \\
&= e^{-\delta t}v'(p_t^h)\sqrt{\Phi(p_t^h)\mu_t^*}\mathrm{d}\bar{B}_t^h,
\end{aligned} \tag{A.6}$$

where in the second line we use the fact that $h_t^*$ is the minimizer of the drift term provided that $\mu_t = \mu_t^*$. Having $v(\cdot)$ satisfying the HJBI equation at $(\mu^*, h^*)$ zero out the drift term above. Using the local-martingale property of $\bar{B}^h$ under $\mathsf{P}^h$, for every finite $t \in \mathbb{R}_+$ we have:

$$\mathsf{E}^h\left[G_t(\mu^*, h)\right] \geq G_0(\mu^*, h) = v(p) \tag{A.7}$$

Note that $G_t(\mu^*, h) \to G_\infty(\mu^*, h) := \delta\int_0^\infty g(\mu_s^*, h_s, p_s^h)\mathrm{d}s$ almost surely. Further, every $h \in \mathcal{H}$ is assumed bounded, hence one can use dominated convergence theorem and let $t \to \infty$ to get:

$$\mathsf{E}^h\left[G_\infty(\mu^*, h)\right] \geq v(p) \tag{A.8}$$

Taking the infimum over all $h \in \mathcal{H}$ leads to:

$$\inf_{h\in\mathcal{H}} \mathsf{E}^h\left[G_\infty(\mu^*, h)\right] \geq v(p) \tag{A.9}$$

Hence,

$$\sup_{\mu\in\mathcal{U}}\inf_{h\in\mathcal{H}} \mathsf{E}^h\left[G_\infty(\mu, h)\right] \geq v(p). \tag{A.10}$$

Now choose an arbitrary $\mu \in \mathcal{U}$ and set $h = \tilde{h} := -\alpha^{-1}\sigma\delta\sqrt{\mu}$:

$$\begin{aligned}
\mathrm{d}G_t(\mu, \tilde{h}) &= \delta e^{-\delta t}\left[g(\mu_t, \tilde{h}_t, p_t^{\tilde{h}}) - v(p_t^{\tilde{h}}) + \frac{\mu_t}{2\delta}\Phi(p_t^{\tilde{h}})v''(p_t^{\tilde{h}})\right]\mathrm{d}t + e^{-\delta t}v'(p_t^{\tilde{h}})\sqrt{\Phi(p_t^{\tilde{h}})\mu_t}\mathrm{d}\bar{B}_t^{\tilde{h}} \\
&\leq \delta e^{-\delta t}v'(p_t^{\tilde{h}})\sqrt{\Phi(p_t^{\tilde{h}})\mu_t}\mathrm{d}\bar{B}_t^{\tilde{h}}
\end{aligned} \tag{A.11}$$

This implies that $\mathsf{E}^{\tilde{h}}\left[G_t(\mu, \tilde{h})\right] \leq v(p)$ for every finite $t \in \mathbb{R}_+$. Using dominated convergence theorem under $\mathsf{P}^{\tilde{h}}$, while sending $t \to \infty$ leads to $\mathsf{E}^{\tilde{h}}\left[G_\infty(\mu, \tilde{h})\right] \leq v(p)$. Now, one can take the infimum over $\mathcal{H}$ and get:

$$\inf_{h\in\mathcal{H}} \mathsf{E}^h\left[G_\infty(\mu, h)\right] \leq v(p) \tag{A.12}$$

This inequality holds for every arbitrary $\mu \in \mathcal{U}$, hence taking the supremum on $\mathcal{U}$ leads to:

$$\sup_{\mu\in\mathcal{U}} \inf_{h\in\mathcal{H}} \mathsf{E}^h \left[G_\infty(\mu, h)\right] \leq v(p) \tag{A.13}$$

Equations (A.10) and (A.13) together imply that $v(p) = \sup_{\mu\in\mathcal{U}} \inf_{h\in\mathcal{H}} \mathsf{E}^h \left[G_\infty(\mu, h)\right]$, that concludes the proof.

### A.4 Proof of theorem 5.1

For the proof of this proposition we need few lemmas.

**Lemma A.1.** *For any $p \in (0, 1)$ the value function is lower bounded by* $\max\left\{r, m(p) - \frac{\sigma^2\delta}{2\alpha}\right\}$.

*Proof.* By replacing nature's best response $h = -\alpha^{-1}\sigma\delta\sqrt{\mu}$ in (3.9), one gets the following payoff representation:

$$v(p) = \sup_{\mu} \mathsf{E}^\mu \left[\delta \int_0^\infty e^{-\delta t} \left((1-\mu_t)r + \mu_t m(p_t^\mu) - \mu_t \frac{\sigma^2\delta}{2\alpha}\right) \mathrm{d}t\right], \tag{A.14}$$

where $\mathsf{E}^\mu$ and $p_t^\mu$ are resp. the probability measure and the posterior probability obtained from $h = -\alpha^{-1}\sigma\delta\sqrt{\mu}$. Furthermore, using the local-martingale property of $m(p_t^h)$, we get the following inequality for every $\mathcal{G}_0$-measurable control process, i.e $\mu_t \in \mathcal{G}_0$ for all $t \in \mathbb{R}_+$, and hence $\mu_t \equiv \mu$ (up to evanescence):

$$\begin{aligned} v(p) &= \sup_{\mu\in\mathcal{U}} \mathsf{E}^\mu \left[\delta \int_0^\infty e^{-\delta t} \left((1-\mu_t)r + \mu_t m(p_t) - \mu_t \frac{\sigma^2\delta}{2\alpha}\right) \mathrm{d}t\right] \\ &\geq \sup_{\mu\in\mathcal{G}_0} \left\{ \mathsf{E}^\mu \left[\delta \int_0^\infty e^{-\delta t} \left((1-\mu)r + \mu m(p_0) - \mu \frac{\sigma^2\delta}{2\alpha}\right) \mathrm{d}t\right] + \mathsf{E}^\mu \left[\delta \int_0^\infty e^{-\delta t} \mu x_t^h \mathrm{d}t\right] \right\} \end{aligned} \tag{A.15}$$

Here $x^h$ is the local-martingale part of $m(p^h)$ resulted from lemma 3.6. Having set $\mu \in \mathcal{G}_0$, the expectation of the second term vanishes due to the $\mathsf{P}^h$-local-martingale property of $x^h$ and using the dominated convergence theorem for approximating the infinite horizon integral with finite counterparts. This proves the lower bound on $v(p)$. □

**Lemma A.2.** *Let $\mathcal{S}_i$ be subset of $[0, 1]$ where the DM optimally chooses the $i$-th project if $p \in \mathcal{S}_i$, where $i \in \{1, 2\}$. Then the value function is convex restricted to each of these subsets.*

*Proof.* On $\mathcal{S}_1$ the value function is identical to $r$, and hence is convex. On $\mathcal{S}_2$ the DM chooses

the second arm and $\mu = 1$, hence

$$v(p) = m(p) - \frac{\sigma^2\delta}{2\alpha} + \frac{1}{2\delta}\Phi(p)v''(p), \tag{A.16}$$

which implies that

$$\frac{1}{2\delta}\Phi(p)v''(p) = v(p) - m(p) + \frac{\sigma^2\delta}{2\alpha} \geq \left(m(p) - \frac{\sigma^2\delta}{2\alpha}\right) - m(p) + \frac{\sigma^2\delta}{2\alpha} = 0. \tag{A.17}$$

Therefore, $v''(p) \geq 0$ and hence the restriction of $v$ onto $\mathcal{S}_2$ is also convex. □

**Lemma A.3.** *The subsets $\mathcal{S}_1$ and $\mathcal{S}_2$ are connected subsets of $[0,1]$.*

*Proof.* First note that $[0,1] = \mathcal{S}_1 \cup \mathcal{S}_2$, therefore the case of one subset being the empty set and the other being the whole unit interval trivially passes the lemma. Now assume both subsets are non-empty, and suppose $\mathcal{S}_1$ is not connected. Therefore, it must contain two disjoint open intervals, say $(a_1, b_1)$ and $(a_2, b_2)$, such that $b_1 < a_2$. This means that $[b_1, a_2] \subset \mathcal{S}_2$. The continuity must holds at the boundaries, namely $v(b_1) = v(a_2) = r$, otherwise there appears an *arbitrage* opportunity for the DM and she could improve her strategy subsets, $\mathcal{S}_1$ and $\mathcal{S}_2$, so as to strictly be better off. Also, one can easily confirm from (A.14) that $v(\cdot)$ is a non-decreasing function in $p$. Since $v(\cdot)$ is always greater than or equal to $r$, then $v \equiv r$ on the entire $[b_1, a_2]$. This means essentially $[b_1, a_2] \subset \mathcal{S}_1$ that violates the initial assumption on $\mathcal{S}_1$. Therefore, $\mathcal{S}_1$ must be a connected subset of $[0,1]$. We use the proof by contradiction again to show $\mathcal{S}_2$ is connected as well. Suppose it is not, then it contains two disjoint open sets, say $(c_1, d_1)$ and $(c_2, d_2)$ such that $d_1 < c_2$. Note that at the boundary points the continuity must hold — precisely to rule out the arbitrage — that means $v(d_1) = v(c_2) = r$. This means either $v \equiv r$ on $(c_1, d_1)$, which then one should include this interval in $\mathcal{S}_1$, or there exists some point $z \in (c_1, d_1)$ such that $v(z) > r$. This violates the non-decreasingness of $v$, and hence concludes the proof. □

The existence of cut-off strategy now falls out of the connectedness of $\mathcal{S}_1$ and $\mathcal{S}_2$ from previous lemma and monotonicity of $v(\cdot)$. It is thus left to prove the global convexity of $v(\cdot)$. For this denote the cut-off point by $\bar{p}$, and note that $\mathcal{S}_1 = [0, \bar{p}]$ and $\mathcal{S}_2 = (\bar{p}, 1]$.[20] So far, we know that $v$ is separately convex on $\mathcal{S}_1$ and $\mathcal{S}_2$. To show that convexity is preserved on the whole region $[0,1]$, we pick the arbitrary points $p_1 \in \mathcal{S}_1$ and $p_2 \in \mathcal{S}_2$ and an arbitrary mixing weight $\xi \in (0,1)$. Define $p_\xi := \xi p_2 + (1-\xi)p_1$. If $p_\xi \in \mathcal{S}_1$, then $\xi v(p_2) + (1-\xi)v(p_1)$

[20] It is not important whether $\bar{p}$ belongs to $\mathcal{S}_1$ or $\mathcal{S}_2$, since essentially the DM is indifferent between two arms when her belief is $\bar{p}$. However, since we laid out the HJB equation on $\mathcal{S}_2$, it is preferred to have an open set as the domain of the differential equation.

is clearly greater than or equal to $v(p_\xi) = r$. Now suppose $p_\xi \in \mathcal{S}_2$, then

$$\xi v(p_2) + (1-\xi)v(p_1) = \xi v(p_2) + (1-\xi)v(\bar{p}) \geq v\left(\xi p_2 + (1-\xi)\bar{p}\right) \geq v(p_\xi), \tag{A.18}$$

where for the first inequality we used the convexity of $v$ on $\mathcal{S}_2$, and for the second one we used the monotonicity of $v$ and the fact that $p_1 \leq \bar{p}$. This concludes the global convexity of $v$, and hence the proof the theorem.

### A.5 Optimal constants for value function with unambiguous information source

The following list is the set of all boundary conditions required for the DM's best-responding:

$$\begin{aligned}
&\text{(value-matching)} && r + c_1\tilde{p}^{\lambda_1}(1-\tilde{p})^{1-\lambda_1} = m(\tilde{p}) - \frac{\sigma^2\delta}{2\alpha} + c_2\tilde{p}^{1-\lambda_2}(1-\tilde{p})^{\lambda_2} \\
&\text{(smooth-pasting)} && c_1\left(\frac{\lambda_1}{\tilde{p}} - \frac{1-\lambda_1}{1-\tilde{p}}\right)\tilde{p}^{\lambda_1}(1-\tilde{p})^{1-\lambda_1} = \left(\overline{\theta} - \underline{\theta}\right) + c_2\left(\frac{1-\lambda_2}{\tilde{p}} - \frac{\lambda_2}{1-\tilde{p}}\right)\tilde{p}^{1-\lambda_2}(1-\tilde{p})^{\lambda_2} \\
&\text{(super-contact)} && \frac{c_1\delta}{\tilde{p}^2(1-\tilde{p})^2\varphi(\gamma)^2}\tilde{p}^{\lambda_1}(1-\tilde{p})^{1-\lambda_1} = \frac{c_2\delta}{\tilde{p}^2(1-\tilde{p})^2\left(\varphi(\sigma)^2 + \varphi(\gamma)^2\right)}\tilde{p}^{1-\lambda_2}(1-\tilde{p})^{\lambda_2}
\end{aligned} \tag{A.19}$$

Therefore, the value of constants $(c_1, c_2)$ are determined in terms of the cut-off point $\tilde{p}$:

$$c_1 = \frac{\frac{\sigma^2}{\gamma^2}\left(r - m(\tilde{p}) + \frac{\sigma^2\delta}{2\alpha}\right)}{\tilde{p}^{\lambda_1}(1-\tilde{p})^{1-\lambda_1}},\ c_2 = \frac{\left(1 + \frac{\sigma^2}{\gamma^2}\right)\left(r - m(\tilde{p}) + \frac{\sigma^2\delta}{2\alpha}\right)}{\tilde{p}^{1-\lambda_2}(1-\tilde{p})^{\lambda_2}} \tag{A.20}$$

Plugging the above two constants into the smooth-pasting conditions leads to (6.7).

### A.6 Proof of proposition 6.2

The associated equations for the cut-off probabilities in each case are expressed in (5.4) and (6.7). First, we show that $\Lambda \geq \lambda$ for any combination of variables. One can easily check from definition of $\Lambda$ and $\lambda$ that $\Lambda \geq \lambda$ iff

$$\frac{\sigma^2}{\gamma^2}\sqrt{1+\beta\gamma^2} + \left(1 + \frac{\sigma^2}{\gamma^2}\right)\sqrt{1 + \beta\frac{\sigma^2\gamma^2}{\sigma^2+\gamma^2}} \geq \sqrt{1+\beta\sigma^2}, \tag{A.21}$$

in that we denote $\beta := 8\delta/(\overline{\theta} - \underline{\theta})^2$. Because $\gamma^2 \geq \sigma^2\gamma^2/(\sigma^2+\gamma^2)$ the *lhs* is larger than

$$\left(1 + \frac{2\sigma^2}{\gamma^2}\right)\sqrt{1 + \beta\frac{\sigma^2\gamma^2}{\sigma^2+\gamma^2}}. \tag{A.22}$$

Therefore, a sufficient condition for (A.21) to hold is $\left(1+\frac{2\sigma^2}{\gamma^2}\right)^2 \geq \frac{\left(1+\frac{\sigma^2}{\gamma^2}\right)\left(1+\beta\sigma^2\right)}{1+\frac{\sigma^2}{\gamma^2}+\beta\sigma^2}$, which holds because

$$\left(1+\frac{2\sigma^2}{\gamma^2}\right) \geq 1 \geq \frac{1+\beta\sigma^2}{1+\frac{\sigma^2}{\gamma^2}+\beta\sigma^2}. \tag{A.23}$$

Now we verify that $\tilde{p} \geq \bar{p}$. For this, note that $\tilde{p} = 0$ if $\Lambda \leq \eta$, in that case $\lambda \leq \eta$ which implies $\bar{p} = 0$. For the region $\lambda > \eta$ both cut-offs are positive. They are equal to one if $\eta \geq 1$ and are strictly smaller than one if $\eta < 1$, in that case $\tilde{p} \geq \bar{p}$ because $\Lambda \geq \lambda$.