Risk aversion and optimal strategies in a one-armed bandit problem: an application to road choice

Jean-Philippe Chancelier†, Michel de Lara†,[*] André de Palma‡

†CERMICS, École nationale des ponts et chaussées, 6 et 8 avenue Blaise Pascal,
Champs sur Marne — 77455 Marne la Vallée Cedex 2

‡Université de Cergy-Pontoise, École nationale des ponts et chaussées, CORE
and Senior Member of Institut universitaire de France

April 1, 2005

Abstract. We study risk aversion with the one-armed bandit problem where (for example) a driver selects, day after day, either a safe or a random road. Four information regimes are envisaged. The visionary driver knows, before hand, with certainty the travel time on the random road, while the locally informed driver needs to select a road to acquire information on it. Two intermediary information regimes (fully and globally) are also envisaged. We analyze these four regimes and compare the optimal strategies and the individual benefits, with respect to individual risk aversion. A numerical example also illustrates the impact of risk aversion on dynamic optimal strategies.

Key words: bandit problem, road choice, risk aversion, uncertainty, travel time, expected utility theory, intelligent transport system
Journal of Economic Literature Classification Number: D830

We would like to thank Nathalie Picard and Maxim Engers for fruitful discussions.

[*] Corresponding author: delara@cermics.enpc.fr

1 Introduction

Choice under uncertainty has a long history, starting with mathematicians, such as Bernoulli, De Moivre, Pascal and Fermat who developed the first steps towards a formal probabilistic decision theory, and with economists such as Dupuis who introduced the concept of utility. These seminal studies have attracted the attention of many famous economists such as Keynes (who wrote his first book on probabilities) or Knight (who discussed the idea of uncertainty and provided the basis for the subjective utility theory). Knight's work was later on pursued by de Finetti and Savage, who started to formalize the ideas of risk and uncertainty. Our paper is based on two streams of research related to risk and uncertainty that we briefly sketch below.

The first stream, the study of choice undertaken by risk averse individuals *facing risky situations*, has been formalized by expected utility theory. During the Second World War, a difficult book was published, by von Neumann and Morgenstern ([15]). These authors, referred to as VNM, provided an original and seminal method to introduce risk in the standard decision theory. Recall here that risk corresponds to a random variable which determines the payoffs and whose distribution is known by the decision maker. Even if it has been challenged on many grounds by experimental economists and psychologists, such as Kahneman and Tversky [10], till now VNM theory has remained the benchmark, not yet successfully challenged (according to many scholars in decision theory).

The second stream of research is related to *uncertainty.* The prototype problem of choice under uncertainty is the multi-armed bandit problem studied independently, as far as we know, also during the Second World War[1]. The key difference is that when the decision maker faces uncertainty, the distribution of the underlying random variable is not known by the decision maker. For example, in the multi-armed bandit, a "job" should be performed iteratively by selecting, time period after time period, one out of n machines. The performances of the machines are unknown, but the decision maker acquires some information on the machine by performing the job with it. Each time a machine is selected, it provides information about its performance, but in a costly manner if the job is poorly done.

Situations of decision under risk provide limiting cases since, usually, individuals do not know the degree of risk they are facing. In general, the level of uncertainty is endogeneous and could be reduced if enough effort were provided by the decision maker. We will show that the level of uncertainty reduction that the decision maker selects is a function of his degree of risk aversion, which provides a link between risk aversion and uncertainty reduction.

In this paper, we wish to consider situations which contain the two facets mentioned above (risk aversion and uncertainty reduction). To fix ideas, we briefly discuss below three examples. Many other examples can be constructed in other fields, including in biology (biological essays and experimental design), and in economics (search models).

1. Consider an individual who uses a known technology and contemplates to shift to a new technology that has been introduced in the market. This new technology is

[1] Before we further discuss this problem, let us hear Professor Whittle, while he was commenting the famous Gittins's paper about the multi-armed bandit problem: "*the problem is a classic one; it was formulated during the war, and efforts to solve it so sapped the energies and minds of Allied analysts that the suggestion was made that the problem be dropped over Germany, as the ultimate instrument of intellectual sabotage*".

unknown, and could be better or worse than the current one for this individual. In order to acquire information about this new technology, the individual has to use it (it is an experience good). Each period of time, the individual makes his mind and either continues to use the current technology, or shift (but not necessarily in an irreversible manner) to the new one. Should the individual try to know more about the new technology, and what is the meaning of technology adoption in this context? We will study how risk aversion modifies individual strategies and payoffs.

2. Consider now a risk averse investor who is ready to invest in two financial instruments: money market and small business. The returns on money market can be assumed to be known (this is realistic in nominal terms), while the investor may have some prior about the returns in small business, which could lead to a higher expected return (risk premium), but may also lead to a zero return (in case of bankrupcy). The investor could initially select the random investment, and stick to it, or, after a learning period, may discover that small business is not that attractive and shift back to the safe alternative. Once he selects the money market investment, if he does not learn anything more about the benefits and the danger of investing in small businesses, he should therefore rationally stick to it.

3. Finally, in our third example, a commuter has access to two modes to go from home to work. Either he can use a very reliable public transportation system (the travel time in this case is given by the schedule) or he could use his car. However, the travel time by car is not guaranteed but it could be assumed stationary. To economize on notations, we will assume that the travel time on the road takes discrete values. These values are known, but the probabilities of occurrence of these values are not known. For example, the travel time could be 20 minutes under normal conditions, but 30 minutes when an accident occurs, which implies that one lane is close. However, the driver does not know the (stationary) frequency of closing. The model in this case will determine how drivers select day after day each alternative.

In this paper, we analyze in depth a slight variant of this last example. In this context, we compare different information regimes and also drivers who differ with respect to their degree of risk aversion. The individual may acquire information either via an exogenous information system or via personal experience. The traffic manager may have access to a large enough set of past occurrence so that he is able to compute an estimate of the probability of good and bad traffic conditions, in the simple case where travel time takes only two values. Alternatively, the traffic manager may not have access to past data, and in this case, the only information that could be transmitted is the realization of good or bad traffic conditions. With limited information, the traffic manager may estimate the probability of occurrence of good and bad traffic conditions, but these estimates depend on a prior about the probability of occurrence. A better alternative is to provide the information about past occurrences, so that each driver may construct his own posterior. For example, carriers sometimes provide the occurrence of delays over a previous period of time and it is up to the user to estimate the probabilities of good or bad state. If there is no information system, the user has only access to the information he has collected. In this case, the random alternative is an experience good, since information about the state of an alternative requires the use of it (here the travel conditions on the road can be observed only if the road is chosen).

We show that the optimal decision problem can be characterized rather easily, except when the only source of information is personal experience. In this case, the choice has a dual effect since it leads to some reward, but it also leads to additional experience, which can be valuable for further choices. Note, however, that this type of dual aspects of choice is not considered in the standard discrete choice theory used in transportation [14, 11].

We show that the road choice between a safe and a random alternative can be formulated as a one-armed bandit problem, but with an extra dimension explored here, the degree of risk aversion. More precisely, we examine in a dynamic context how risk aversion modifies individual decisions under uncertainty. Note that, in our analysis, we do not need to resort to assumptions with respect to aversion to ambiguity (see also [6]). Here we consider that individuals have different attitudes towards risk, but we assume nothing with respect to their attitude towards uncertainty. Such behavior will be an outcome of our rational model. The optimal solution of this problem shows that the individuals have specific propensities to actively and costly reduce the level of uncertainty, according to their degree of risk aversion.

In Section 2, we show how risk aversion can parametrize optimal dynamic strategies. In Section 3, we introduce the model and four information regimes. while in Section 4 we present the corresponding four optimal strategies. In Section 5, we concentrate our attention on situations where users can only reduce incertainty by personnal experience and compare the different solutions. Numerical results are presented in Section 6. Concluding comments and future research directions are briefly provided in the last section. General mathematical proofs are relegated to the Appendix.

2 Risk aversion and the one-armed bandit problem

2.1 Static decision under uncertainty

Consider a decision-maker facing two alternatives S (safe) and R (random). Alternative S yields a deterministic payoff x_S, while alternative R yields a random payoff X whose distribution law ν on $\mathbb{R}$ is not known to the decision-maker. The preferences of the decision-maker are characterized by a (strictly increasing[2] and concave) utility function U.

In this static framework, one needs an additional behavioral assumption to solve this binary choice problem. For example, a standart assumption is that the decision-maker selects the alternative which maximizes the worst expectation under a family of distribution laws on $\mathbb{R}$ (max-min strategies). Alternatively, the decision-maker is assumed to have a prior on ν (a distribution law π_0 on the space of distribution laws on $\mathbb{R}$) and to maximize a (doubly) expected utility. Note that, in this case, only the known mathematical expectation of the prior matters and the initial choice under uncertainty is formally turned into choice under risk.

[2] By a *stricly increasing* function f, we mean that $x > y \Rightarrow f(x) > f(y)$. We reserve the term of *increasing* for functions f such that $x > y \Rightarrow f(x) \geq f(y)$.

2.2 Dynamic decision with costly learning

In a dynamic framework, the decision-maker may learn about ν by observing i.i.d. realizations (X_t) of the random payoff X and update his prior on ν each time he selects the alternative R. The decision-maker selects a $v_t \in \{R, S\}$ at every period t, which gives a stochastic discounted intertemporal utility $J(v(\cdot)) = \sum_{t=0}^{+\infty} \rho^t U(\Phi(v_t, X_{t+1}))$, where $\Phi(S, X_{t+1}) = x_S$, $\Phi(R, X_{t+1}) = X_{t+1}$ and where $\rho \in [0, 1[$ is the discount rate. Given a prior π_0 on ν, one thus obtains a known distribution over the random sequence $(X_t)_{t\geq 1}$, under which the decision-maker maximizes the mathematical expectation of $J(v(\cdot))$.

2.3 Relation with the "classical bandit problem"

The above problem is a case of so called "classical bandit problem", where the state of arm S is degenerate and the state of arm R is the belief $\widehat{\pi}_t$ of the decison-maker regarding the "true" distribution ν. This is a *one*-armed bandit problem because the state of arm S is deterministic stationary and returns a reward $\Psi_S = x_S$. The other arm state $\widehat{\pi}_t$ forms a Markov process whose transitions correspond to the Bayesian updating with respect to the observation of X_t. The reward is $\Psi_R(\pi) = \int \pi(d\nu) \int \nu(d\omega) U(X(\omega))$. Our presentation of bandit problems is quite sketchy, and we send the reader to specialized references such as [7, 8, 16, 1].

The "classical bandit problem" is an example of stochastic control problem with partial information. Such problems are generally turned into problem with full information by introducing conditional laws as a new state as above (see details in [3]). Their solutions are then given by stochastic dynamic programming. However, bandit problems have special solutions, called index strategies, as a consequence of their specific structure: arms are independent and their state only evolves when they are selected, and the decision-maker maximizes a discounted intertemporal utility.

2.4 Index strategies

The "classical bandit problem" refers to a situation of partial information. However, once one chooses as "new state" the conditional law of the "old state", the classical bandit problem becomes a simple "bandit problem", that is one with full information.

Bandit problems with geometric discounting are well known (see [7, 8, 16, 1]) for having optimal strategies that may be characterized by the so called *Gittins indices* (or dynamic allocation indices). To each arm is associated a state and an index function depending upon this state. At a given stage, one compares the values of the different indexes and the optimal strategy of the decision-maker is to select the arm with the higher index. The selected state evolves according to a given transition kernel while the other states remain fixed, and the value of the index of the selected arm is updated in consequence. The process goes on at the next time period.

Consider an arm with state space $\mathbb{Z}$ and reward $\Psi : \mathbb{Z} \to \mathbb{R}$. Starting from state $z_0 \in \mathbb{Z}$ and always selecting this arm yields a stochastic process $(z_t)_{t\geq 0}$ in $\mathbb{Z}$ determined by the given transition kernel. The Gittins index of this arm is the following supremum over stopping times $\tau > 0$ (see [7]):

$$\mu(z) = \sup_{\tau>0} \frac{\mathbb{E}[\sum_{t=0}^{\tau-1} \rho^t \Psi(z_t) \mid z_0 = z]}{\mathbb{E}[\sum_{t=0}^{\tau-1} \rho^t \mid z_0 = z]} \,. \tag{1}$$

This index may be regarded as an average reward over a stopping time, with the rule determining it chosen to maximize this average. Another interpretation is as follows. Allow the additional option of retirement for a lump reward of L. Then, the index is the value of L whicj makes the option of continuation or of retirement (with reward L) equally attractive [16]. In our one-armed bandit problem, the index of arm S is constant ($\mu_S = \Psi_S = U(x_S)$), while $\mu_R(\pi)$ has no analytical expression. Knowing μ_S, μ_R and the state $z_t = \widehat{\pi}_t$ at time t, the optimal strategy is as follows: select the arm with the higher index, that is arm S if $\mu_S \geq \mu_R(\widehat{\pi}_t)$ and arm R else.

2.5 Risk aversion and optimal strategies

We wish to examine how individual risk aversion modifies dynamics of optimal decisions. Recall the Arrow-Pratt definition of absolute risk aversion [13, 9, 5]. By definition, decision-maker with utility function U^M is more risk averse than decision-maker with utility function U^L if U^M is a concave transformation of U^L. Notice that the transformation is necessary strictly increasing because U^M and U^L are strictly increasing.

Proposition 1

Consider two decision-makers with common prior belief π_0 and one more risk averse than the other. Assume that, at the beginning, the more risk averse decision-maker selects arm R based on π_0. Then, so does the less risk averse decision-maker and, as long as the more risk averse decision-maker selects arm R, so does also the less risk averse decision-maker.

Proof. Assume that decision-maker with utility function U^M is more risk averse than decision-maker with utility function U^L. There exists a concave strictly increasing function φ such that $\varphi \circ U^L = U^M$. The state space is here $\mathbb{Z} = \mathcal{P}(\mathcal{P}(\mathbb{R}))$, the space of probabilities on the space of probabilities on $\mathbb{R}$, and the rewards are given by

$$\Psi_S^{M,L} = U^{M,L}(x_S) \quad \text{and} \quad \Psi_R^{M,L}(\pi) = \int \pi(d\nu) \int \nu(d\omega) U^{M,L}(X(\omega)) \,, \quad \forall \pi \in \mathcal{P}(\mathcal{P}(\mathbb{R})) \,. \tag{2}$$

We have $\Psi_S^M = U^M(x_S) = \varphi(U^L(x_S)) = \varphi(\Psi_S^L)$. On the other hand, we have:

$$\begin{aligned}
\Psi_R^M(\pi) &= \int \pi(d\nu) \int \nu(d\omega) U^M(X(\omega)) \\
&= \int \pi(d\nu) \int \nu(d\omega) \varphi(U^L(X(\omega))) \ \text{ since } U^M = \varphi \circ U^L \\
&\leq \varphi(\int \pi(d\nu) \int \nu(d\omega) U^L(X(\omega))) \ \text{ since } \varphi \text{ is concave} \\
&= \varphi(\Psi_R^L(\pi)) \,.
\end{aligned}$$

The end of the proof follows with Proposition 10 in Appendix. □

This Proposition implies that the decision-makers can be ranked by their degree of risk aversion. The most risk averse decision-makers eventually select the safe road, while the least risk averse decision-makers select the random road and stick to it. This simple manner to sort decision-makers will play an essential role when congestion will be taken into account.

A direct consequence of the above Proposition is the following Corollary.

Corollary 2
The mean time spent selecting arm R decreases with the degree of absolute risk aversion.

3 Road choice problem statement

In this section, we describe an elementary road choice problem, with one safe road with known deterministic travel time, and one random in the sense that the travel time is stochastic. For the sake of simplicity, we assume that this latter random travel time may take only a finite number n of known values. Our results would remain valid (but at the price of high technicality) if random travel time were a general random variable. We shall describe different decision problems according to the information available to a driver.

3.1 Basic notations

Consider one origin, one destination and two roads in parallel. At every period $[t, t+1[$ starting at t (for instance a day), denoted by period t in the sequel, the driver selects one of the two roads (safe and random) characterized as follows:

1. The *safe road* S has constant known travel time x_S.
2. The *random road* R has a random travel time X_{t+1} realized at the end of period t. X_{t+1} takes $n \geq 2$ values in $\{x_1, \dots, x_n\}$.

We suppose that x_i occurs with probability $\overline{p}_i \in]0,1[$, $i = 1, \dots, n$, with $\overline{p}_1 + \cdots + \overline{p}_n = 1$. We assume that $x_1, \dots, x_n$ are known to the driver and that

$$x_1 < \cdots < x_n \quad \text{and} \quad x_1 < x_S < x_n \,. \tag{3}$$

An adequate sample space (states of nature) may be $\Omega = \{x_1, \dots, x_n\}^{\mathbb{N}^*}$ with the σ-field $\mathcal{F} = 2^{\Omega}$ of all subsets. The coordinates $X_t(\omega) = \omega(t)$, $t \in \mathbb{N}^*$, form the sequence of random travel times on the random road. Let S_{n-1} denote the simplex of dimension n:

$$S_{n-1} \stackrel{\text{def}}{=} \{(p_1, \dots, p_n) \in \mathbb{R}_+^n \,, \quad p_1 + \cdots + p_n = 1\} \,. \tag{4}$$

We define

$$\overline{p} \stackrel{\text{def}}{=} (\overline{p}_1, \dots, \overline{p}_n) \in S_{n-1} \tag{5}$$

(in fact, $\overline{p}$ belongs to the interior of S_{n-1} by our assumption that $\overline{p}_i \in]0,1[$, $i = 1, \dots, n$) and the *probability* $\mathbb{P}^{\overline{p}}$ on Ω by the marginals, for any $(z_1, \dots, z_t) \in \{x_1, \dots, x_n\}^t$

$$\mathbb{P}^{\overline{p}}(X_1 = z_1, \dots, X_t = z_t) = \prod_{s=1}^{t} [\overline{p}_1 \mathbf{1}_{\{z_t = x_1\}} + \cdots + \overline{p}_n \mathbf{1}_{\{z_t = x_n\}}] \,. \tag{6}$$

The *expectation* under the probability $\mathbb{P}^{\overline{p}}$ is denoted by $\mathbb{E}^{\overline{p}}$.

The *decision* $v_t \in \{S, R\}$ is the road chosen at the beginning of period t. The observation at the end of period t depends on the information regimes envisaged (see Paragraph 3.3). Information could be acquired either by direct observation or *via* some driver information system which may forecast future travel conditions.

We shall denote by Y_{t+1} the travel time experienced by the driver at the end of period t. It is either x_S if he selects the safe road or X_{t+1}. Thus, the *experienced travel time* Y_{t+1} which depends upon both the decision v_t and X_{t+1} is given by the formula

$$Y_{t+1} = \Phi(v_t, X_{t+1}) \,, \tag{7}$$

where the function $\Phi(\cdot,\cdot)$ defined on $\{R, S\} \times \{x_1, \ldots, x_n\}$ is given by

$$\Phi(S, x_1) = \cdots = \Phi(S, x_n) = x_S \quad \text{and} \quad \Phi(R, x_1) = x_1 \,, \ldots \,, \Phi(R, x_n) = x_n \,. \tag{8}$$

3.2 Preference model

The preferences of a driver are characterized by a utility function V and by the discount rate $\rho \in [0, 1[$. We shall call it *driver* $[V, \rho]$. The utility function V is strictly decreasing[3] concave, so that

$$V(x_1) > \cdots > V(x_n) \quad \text{and} \quad V(x_1) > V(x_S) > V(x_n) \,. \tag{9}$$

The rational driver selects a road $v_t \in \{S, R\}$ at every period $t \geq 0$. The nature of the information available to the driver when he selects one of the roads is crucial: we shall develop this point below (Paragraph 3.3). To a sequence $v(\cdot) = (v_0, v_1, \ldots)$ of decisions is associated a stochastic intertemporal utility $\sum_{t=0}^{+\infty} \rho^t V(Y_{t+1})$, where Y_{t+1} depends upon v_t by Eq. (7).

Let the *reward* $G(v, x)$ be defined by the instantaneous utility resulting from road choice and experienced travel time:

$$G(v, x) \overset{\text{def}}{=} V(\Phi(v, x)) \,, \quad \forall v \in \{R, S\} \,, \quad \forall x \in \{x_1, \ldots, x_n\} \,. \tag{10}$$

By Eq. (7), we have $V(Y_{t+1}) = G(v_t, X_{t+1})$. For a sequence $v(\cdot) = (v_0, v_1, \ldots)$ of decisions, the *discounted reward* $J(v(\cdot))$ is the random variable

$$J(v(\cdot)) \overset{\text{def}}{=} \sum_{t=0}^{+\infty} \rho^t G(v_t, X_{t+1}) = \sum_{t=0}^{+\infty} \rho^t V(\Phi(v_t, X_{t+1})) = \sum_{t=0}^{+\infty} \rho^t V(Y_{t+1}) \,. \tag{11}$$

Optimal strategies are given by the maximization of the mathematical expectation (under probability laws specified later) of this random discounted reward.

[3] This is because, in the transportation context, V is decreasing in its argument, the travel time.

3.3 Information regimes

We recall that x_S, x_1, ..., x_n are known to the driver. We shall focus on four information regimes and associated optimization problems with increasing difficulty.

1. At the beginning of every period t, *i.e.* before he makes his decision, the *visionary driver* (v) knows with certainty the travel time on the random road at the end of period t.
2. The *fully informed driver* (ϕ) knows $\overline{p} = (\overline{p}_1, \dots, \overline{p}_n)$;
3. The *globally informed driver* (γ) does not know $\overline{p}$ (but has a prior π_0 on $\overline{p}$) and, at the beginning of period t, knows all past random travel times X_1,..., X_t unconditional on road choice.
4. The *locally informed driver* (λ) does not know $\overline{p}$ (but has a prior π_0 on $\overline{p}$) and, at the beginning of period t, knows only Y_1,..., Y_t given by Eq. (7), that is only past random travel times when he has selected the random road.

In the first and second cases, the optimization problem is easily solved. In the third case, the posterior distribution on $\overline{p}$ is revised at each period, unconditionaly on road choice, and we give the optimal rule. In the fourth case, the posterior distribution on $\overline{p}$ depends on previous decisions. We use Bayesian update rules. This is an illustration of the dual effect of a decision which contributes both to improve the expected intertemporal utility and to provide valuable information for future decisions. This latter case is naturaly formulated as a one-armed bandit problem, and we shall give some insights on its solution.

4 Optimal strategies and information regimes

We provide below the optimal strategies for the four information regimes envisaged. Information regimes are ranked as follows: the visionary driver (Paragraph 4.1) is the best informed driver, while the locally informed driver (Paragraph 4.5) is the most poorly informed one. Intermediary information regimes are considered in Paragraphs 4.2 and 4.4. For the first three regimes, information is independent of the action (road choice) while the last regime requires an active move from the driver to acquire information. As we will show, the last information regime is far more complex to study.

Without loss of generality, we assume that in case of tight the driver selects the safe road.

4.1 The visionary driver

At the beginning of every period t, the *visionary driver* knows X_{t+1}, *i.e.* gets information about the travel time on the random road at the end of period t. Let strategy $v^v(\cdot)$ consists in maximizing each $G(v_t, X_{t+1})$ since v_t may depend upon X_{t+1}. Obviously, the driver selects the random road if $X_{t+1} < x_S$, and the safe one otherwise. Since $G(v_t^v, X_{t+1}) \geq G(v_t, X_{t+1})$

for any v_t, we clearly have the following inequality between random variables (that is for all sample realizations!):

$$J(v^v(\cdot)) = \sum_{t=0}^{+\infty} \rho^t G(v_t^v, X_{t+1}) \geq J(v(\cdot)) = \sum_{t=0}^{+\infty} \rho^t G(v_t, X_{t+1}) \, . \tag{12}$$

It is clearly optimal (stronger than in the mean sense under any probability) to select the random road at the beginning of period t if the future travel time is known to be strictly lower than x_S, and the safe road if not. Thus, the visionary driver does better than any other driver, let it be a causal one (that is one who knows no more than X_1,..., X_t) or even a fellow visionary who might know more than X_{t+1}, since the above inequality states that

$$J(v^v(\cdot)) = \sup_{v(\cdot)} J(v(\cdot)) \, . \tag{13}$$

Notice that the optimal strategy v^v does not depend upon risk aversion. This is not true for the three subsequent regimes.

4.2 The fully informed driver

Recall that the fully informed driver knows $\overline{p}$. Thus, he looks for a strategy $v^\phi(\cdot) = (v_0^\phi, v_1^\phi, \ldots)$ which maximizes the expectation of $J(v(\cdot))$ under probability $\mathbb{P}^{\overline{p}}$:

$$\mathbb{E}^{\overline{p}}[J(v^\phi(\cdot))] = \sup_{v(\cdot)} \mathbb{E}^{\overline{p}}[J(v(\cdot))] \, . \tag{14}$$

Let the relevant road be the optimal road given the knowledge of $\overline{p}$, defined as follows.

Definition 3
For driver $[V, \rho]$, the relevant road *is defined as the random road R if $V(x_S) < \overline{p}_1 V(x_1) + \cdots + \overline{p}_n V(x_n)$ and as the safe road S otherwise.*

For a risk-neutral driver (linear decreasing utility function), the relevant road is the random road R if and only if $x_S > \overline{p}_1 x_1 + \cdots + \overline{p}_n x_n$. The relevant road depends upon specifications on V and $\overline{p}$, and not upon our general assumptions. From now, we exclude the singular case where $V(x_S) = \overline{p}_1 V(x_1) + \cdots + \overline{p}_n V(x_n)$, and we shall assume in the sequel that

$$V(x_S) \neq \overline{p}_1 V(x_1) + \cdots + \overline{p}_n V(x_n) \, . \tag{15}$$

It is clear that the optimal strategy v^ϕ is to select the relevant road. The optimal expected discounted reward is

$$\mathbb{E}^{\overline{p}}[J(v^\phi(\cdot))] = \frac{1}{1-\rho} \max\{V(x_S), \overline{p}_1 V(x_1) + \cdots + \overline{p}_n V(x_n)\} \, . \tag{16}$$

Note that the optimal reward $\mathbb{E}^{\overline{p}}[J(v^\phi(\cdot))]$ of the fully informed driver is then smaller than the optimal reward $\mathbb{E}^{\overline{p}}[J(v^v(\cdot))]$ of the visionary driver (in fact, strictly smaller since $0 < \overline{p}_i < 1$ and by Eq. (9)).

4.3 A common framework for globally and locally informed drivers

Both the globally and locally informed drivers cannot evaluate an expected discounted reward like $\mathbb{E}^{\overline{p}}[J(v(\cdot))]$ since they do not know $\overline{p}$. However, both have a prior law π_0 over $\overline{p}$: π_0 is a distribution over the simplex S_{n-1}. With π_0, we may define the probability $\mathbb{P}^{\pi_0}$ on Ω by the marginals

$$\mathbb{P}^{\pi_0}(X_1 = z_1, \ldots, X_t = z_t) = \int_{S_{n-1}} \pi_0(dp_1 \cdots dp_n) \prod_{s=1}^{t} [p_1 \mathbf{1}_{\{z_t = x_1\}} + \cdots + p_n \mathbf{1}_{\{z_t = x_n\}}] . \tag{17}$$

The above formula must be understood as the integral of the function $(p_1, \ldots, p_n) \hookrightarrow \prod_{s=1}^{t} [p_1 \mathbf{1}_{\{z_t = x_1\}} + \cdots + p_n \mathbf{1}_{\{z_t = x_n\}}]$, defined on the simplex S_{n-1} against the distribution π_0. In particular, with our notation, the integral element $\pi_0(dp_1 \cdots dp_n)$ has to be understood as one over the simplex S_{n-1} and not over $\mathbb{R}^n$. $\mathbb{E}^{\pi_0}$ denotes the expectation under the probability $\mathbb{P}^{\pi_0}$. When $\pi_0 = \delta_p$, we simplify $\mathbb{P}^p \overset{\text{def}}{=} \mathbb{P}^{\delta_p}$ (this is coherent with the definition Eq. (6) of $\mathbb{P}^{\overline{p}}$).

A formulation with a state space of probabilities on S_{n-1}

To solve such problems where the element $\overline{p}$ of the simplex S_n is unknown, it is classical to introduce the space $\mathcal{P}(S_{n-1})$ of probabilities on S_{n-1} as the state space. For this, let us first introduce some notations. Let $[\pi]_i$ denote, for $i = 1, \ldots, n$,

$$\forall \pi \in \mathcal{P}(S_{n-1}) , \quad [\pi]_i \overset{\text{def}}{=} \int_{S_{n-1}} p_i \pi_0(dp_1 \cdots dp_n) , \tag{18}$$

and

$$[\pi] \overset{\text{def}}{=} ([\pi]_1, \ldots, [\pi]_n) \in S_{n-1} , \tag{19}$$

For all $v \in \{R, S\}$ and $\pi \in \mathcal{P}(S_{n-1})$, let us define a new reward $\widetilde{G}(v, \pi)$ by

$$\widetilde{G}(v, \pi) \overset{\text{def}}{=} [\pi]_1 G(v, x_1) + \cdots + [\pi]_n G(v, x_n) . \tag{20}$$

The reward for the safe road S is $\widetilde{G}(S, \pi) = V(x_S)$, while for the random road R, it is $\widetilde{G}(R, \pi) = [\pi]_1 V(x_1) + \cdots + [\pi]_n V(x_n)$.

For $i = 1, \ldots, n$, let M_t^i be one plus the number of periods in which x_i has been realized:

$$M_t^i \overset{\text{def}}{=} 1 + \sum_{s=1}^{t} \mathbf{1}_{\{X_t = x_i\}} . \tag{21}$$

We have $(M_t^1 - 1) + \cdots + (M_t^n - 1) = t$. Let $\widehat{\pi}_t^\gamma \in \mathcal{P}(S_{n-1})$ be defined by:

$$\widehat{\pi}_t^\gamma(dp_1 \cdots dp_n) \overset{\text{def}}{=} \frac{\pi_0(dp_1 \cdots dp_n) p_1^{M_t^1 - 1} \cdots p_n^{M_t^n - 1}}{\int_{S_{n-1}} \pi_0(dp_1 \cdots dp_n) p_1^{M_t^1 - 1} \cdots p_n^{M_t^n - 1}} . \tag{22}$$

Note that $\widehat{\pi}_t^\gamma$ may be computed by induction:

$$\widehat{\pi}_{t+1}^\gamma(dp_1 \cdots dp_n) = \frac{\widehat{\pi}_t^\gamma(dp_1 \cdots dp_n)[p_1 \mathbf{1}_{\{X_{t+1} = x_1\}} + \cdots + p_n \mathbf{1}_{\{X_{t+1} = x_n\}}]}{\int_{S_{n-1}} \widehat{\pi}_t^\gamma(dp_1 \cdots dp_n)[p_1 \mathbf{1}_{\{X_{t+1} = x_1\}} + \cdots + p_n \mathbf{1}_{\{X_{t+1} = x_n\}}]} . \tag{23}$$

The case of β priors

Recall that a beta law $\beta(r,s)$ is defined on $[0,1]$ by (see [12])

$$\beta(r,s) \overset{\text{def}}{=} \frac{1}{B(r,s)} p^{r-1}(1-p)^{s-1}\mathbf{1}_{[0,1]}(p) \quad \text{where} \quad B(r,s) = \int_0^1 p^{r-1}(1-p)^{s-1}dp \,. \tag{24}$$

This corresponds to the case $n = 2$. This definition may easily be extended to distributions over the simplex S_{n-1}. By the beta law $\beta(m_1, \ldots, m_n)$ on S_{n-1}, we mean

$$\beta(m_1, \ldots, m_n) \overset{\text{def}}{=} \frac{\varsigma_{n-1}(dp_1 \cdots dp_n) p_1^{m_1-1} \cdots p_n^{m_n-1}}{\int_{S_{n-1}} \varsigma_{n-1}(dp_1 \cdots dp_n) p_1^{m_1-1} \cdots p_n^{m_n-1}} \tag{25}$$

where $\varsigma_{n-1}(dp_1 \cdots dp_n)$ is the uniform distribution over the simplex S_{n-1}. A computation by induction on n gives

$$\forall i = 1, \ldots, n \,, \quad [\beta(m_1, \ldots, m_n)]_i = \frac{m_i}{m_1 + \cdots + m_n} \,. \tag{26}$$

When $\pi_0 = \beta(m_1^0, \ldots, m_n^0)$ (in particular the uniform law on S_{n-1} for $m_1^0 = \cdots = m_n^0 = 1$), then $\widehat{\pi}_t^\gamma = \beta(m_1^1 + M_t^1, \ldots, m_n^0 + M_t^n)$ and

$$\forall i = 1, \ldots, n \,, \quad [\widehat{\pi}_t^\gamma]_i = \frac{m_i^0 + M_t^i}{m_1^0 + M_t^1 + \cdots + m_n^0 + M_t^n} \,. \tag{27}$$

4.4 The globally informed driver

Problem statement

The information available to the globally informed driver is the so called *history* $\mathcal{X}_t$ up to time t: $\mathcal{X}_t$ is the σ-field $\mathcal{X}_t \overset{\text{def}}{=} \sigma(X_1, \ldots, X_t)$, that is all past travel times on the random road. For every period t, the decision v_t is measurable with respect to history $\mathcal{X}_t$: we shall denote this by $v_t \preceq \mathcal{X}_t$.

Thus, the globally informed driver looks for a strategy $v^\gamma(\cdot) = (v_0^\gamma, v_1^\gamma, \ldots)$ to maximize the expectation of $J(v(\cdot))$ under probability $\mathbb{P}^{\pi_0}$, where the decision v_t at the beginning of period $[t, t+1[$ depend upon the information $\mathcal{X}_t$ available at this time:

$$\mathbb{E}^{\pi_0}[J(v^\gamma(\cdot))] = \sup_{v_t \preceq \mathcal{X}_t, t \geq 0} \mathbb{E}^{\pi_0}[J(v(\cdot))] \,. \tag{28}$$

Optimal strategies

Proposition 4
The optimal globally informed driver

1. *Selects the safe road if and only if*

$$[\widehat{\pi}_t^\gamma]_1 V(x_1) + \cdots + [\widehat{\pi}_t^\gamma]_n V(x_n) \leq V(x_S) \,. \tag{29}$$

2. *Always selects the relevant road after a random number of periods, if its prior* π_0 *is a beta law.*

Proof.

1. By Eq. (17), we may establish that, for any $\pi_0 \in \mathcal{P}(S_{n-1})$:
$$\forall i = 1, \dots, n\,, \quad \mathbb{P}^{\pi_0}(X_{t+1} = x_i \mid \mathcal{X}_t) = [\widehat{\pi}_t^\gamma]_i\,. \tag{30}$$
Thus, by Lemma 11 in the Appendix, we may write
$$\mathbb{E}^{\pi_0}[J(v(\cdot))] = \mathbb{E}^{\pi_0}[\sum_{t=0}^{+\infty} \rho^t \widetilde{G}(v_t, \widehat{\pi}_t^\gamma)]\,. \tag{31}$$
This property, together with Eq. (23), turns the original problem Eq. (28) into an optimal stochastic control problem with state $\widehat{\pi}_t^\gamma$. This state is interpreted as the posterior law of $\overline{p}$ knowing history $\mathcal{X}_t$.
We have
$$\begin{aligned} \sup_{v_t \preceq \mathcal{X}_t, t \geq 0} \mathbb{E}^{\pi_0}[J(v(\cdot))] &= \sup_{v_t \preceq \mathcal{X}_t, t \geq 0} \sum_{t=0}^{+\infty} \rho^t \mathbb{E}^{\pi_0}[\widetilde{G}(v_t, \widehat{\pi}_t^\gamma)] \quad \text{by Eq. (31)} \\ &= \sum_{t=0}^{+\infty} \rho^t \sup_{v_t \preceq \mathcal{X}_t} \mathbb{E}^{\pi_0}[\widetilde{G}(v_t, \widehat{\pi}_t^\gamma)]\,, \end{aligned}$$
since $\widehat{\pi}_t^\gamma$ depends upon $\mathcal{X}_t$, and this latter does not depend upon any control v_s. Thus, the optimization problem has now become a sequence of distinct optimization problems $\sup_{v_t \preceq \mathcal{X}_t} \mathbb{E}^{\pi_0}[\widetilde{G}(v_t, \widehat{\pi}_t^\gamma)]$. Since $\widehat{\pi}_t^\gamma$ is $\mathcal{X}_t$-measurable, we have
$$\sup_{v_t \preceq \mathcal{X}_t} \mathbb{E}^{\pi_0}[\widetilde{G}(v_t, \widehat{\pi}_t^\gamma)] = \mathbb{E}^{\pi_0}[\sup_{v_t \in \{R,S\}} \widetilde{G}(v_t, \widehat{\pi}_t^\gamma)]\,. \tag{32}$$
Thus, the optimal strategy $v_t(\cdot)$ is given by
$$\forall t \geq 0\,, \quad v_t^\gamma = \arg\max_{v \in \{R,S\}} \widetilde{G}(v, \widehat{\pi}_t^\gamma)\,. \tag{33}$$
With Eq. (20), this gives Eq. (29). In other words, denoting
$$\Gamma(p_1, \dots, p_n) \overset{\text{def}}{=} p_1 V(x_1) + \cdots + p_n V(x_n) - V(x_S) \tag{34}$$
the driver selects road S if and only if $\Gamma([\widehat{\pi}_t^\gamma]) \leq 0$.

2. Under probability $\mathbb{P}_{\overline{p}}$, $(X_t)_{t \geq 1}$ are i.i.d. random variables and the law of large number gives
$$\forall i = 1, \dots, n\,, \quad \frac{M_t^i}{t} \to_{t \to +\infty} \overline{p}_i\,, \quad \mathbb{P}_{\overline{p}} \quad \text{a.s.}$$
Thus by Eq. (27), we have that :
$$[\widehat{\pi}_t^\gamma] \to \overline{p}\,, \quad \mathbb{P}_{\overline{p}} \text{ a.s.} \tag{35}$$
As a consequence
$$\Gamma([\widehat{\pi}_t^\gamma]) \to \Gamma(\overline{p})\,, \quad \mathbb{P}_{\overline{p}} \quad \text{a.s.}$$
Having excluded the singular case where $\Gamma(\overline{p}) = 0$, we have that, for t large enough, $\Gamma([\widehat{\pi}_t^\gamma])$ has the same sign than $\Gamma(\overline{p})$. Thus, the globally informed driver selects the relevant road after a random number of periods.

□

The proof would still hold if π_0 where a convex combination of beta laws.

4.5 The locally informed driver as a one-armed bandit problem

The locally informed driver has the same information as the globally informed driver up to the point where the former leaves the random road.

Problem statement

The information available to the locally informed driver consists only of experienced travel times up to period t. We represent this by the σ-field

$$\mathcal{Y}_t \overset{\text{def}}{=} \sigma(Y_1, \dots, Y_t) = \sigma(\Phi(v_0, X_1), \dots, \Phi(v_{t-1}, X_t)) \,, \tag{36}$$

where Φ is defined in Eq. (8). The decision v_t is measurable with respect to information $\mathcal{Y}_t$ ($v_t \preceq \mathcal{Y}_t$). The difficulty comes from the fact that now $\mathcal{Y}_t$ depends upon past v_0,..., v_{t-1} as may be seen in Eq. (36).

For the locally informed driver, the prior law π_0 differs from $\delta_{\overline{p}}$ since $\overline{p}$ is not known. The optimization problem

$$\mathbb{E}^{\pi_0}[J(v^\gamma(\cdot))] = \sup_{v_t \preceq \mathcal{Y}_t, t \geq 0} \mathbb{E}^{\pi_0}[J(v(\cdot))]$$

is classically formulated as an armed-bandit problem as discussed below.

Formulation as a one arm bandit problem

For $i = 1, \dots, n$, let us define N_t^i (one plus the number of times x_i has been observed) by

$$N_t^i \overset{\text{def}}{=} 1 + \sum_{s=1}^{t} \mathbf{1}_{\{X_t = x_i\}} \,. \tag{37}$$

Note that $(N_t^1 - 1) + \cdots + (N_t^n - 1) \leq t$, but not necessarily equal to t. Let us also define a distribution $\widehat{\pi}_t^\lambda$ on S_{n-1} by

$$\widehat{\pi}_t^\gamma(dp_1 \cdots dp_n) \overset{\text{def}}{=} \frac{\pi_0(dp_1 \cdots dp_n) p_1^{N_t^1 - 1} \cdots p_n^{N_t^n - 1}}{\int_{S_{n-1}} \pi_0(dp_1 \cdots dp_n) p_1^{N_t^1 - 1} \cdots p_n^{N_t^n - 1}} \,. \tag{38}$$

As for the globally informed driver, and for the same reasons, the state is here $\widehat{\pi}_t^\lambda$, interpreted as the posterior law of $\overline{p}$ knowing history $\mathcal{Y}_t$.

Now, observe that the state $\widehat{\pi}_t^\lambda$ varies only when the random road is selected: this characterizes bandit problems where one job evolves only if selected.

Gittins indexes

The locally informed driver strategies are expressed by means of $\widehat{\pi}_t^\lambda$ and of the so called Gittins indexes (see [7] and definition Eq. (1)) μ_S and μ_R given below. When $\mu_S \geq \mu_R(\widehat{\pi}_t^\lambda)$, the locally informed driver selects the safe road at period t, and conversely.

The index μ_S of the safe road is the constant reward $\widetilde{G}(S, \widehat{\pi}_t^\lambda) = V(x_S)$. Indeed, by Eq. (1) with state space $\mathbb{Z} = \mathcal{P}(S_{n-1})$ and a constant reward, we have that the index μ_S is constant:

$$\forall \pi \in \mathcal{P}(S_{n-1}) \, , \quad \mu_S(\pi) = V(x_S) \, . \tag{39}$$

The index μ_R of the random road is the following supremum over stopping times $\tau > 0$:

$$\mu_R(\pi) = \sup_{\tau > 0} \frac{\mathbb{E}^\pi [\sum_{t=0}^{\tau-1} \rho^t \widetilde{G}(R, \widehat{\pi}_t^\gamma)]}{\mathbb{E}^\pi [\sum_{t=0}^{\tau-1} \rho^t]} \, . \tag{40}$$

Note that $\widehat{\pi}_t^\gamma$, defined by Eq. (22) with $\pi_0 = \pi$, appears in the above formula. Indeed, when the random road is always selected, then $M_t^i = N_t^i$ for $i = 1, \ldots, n$ and the state $\widehat{\pi}_t^\lambda$ coincides with $\widehat{\pi}_t^\gamma$.

By taking the specific stopping time $\tau = 1$ in (40), we obtain the inequality

$$\forall \pi \in \mathcal{P}(S_{n-1}) \, , \quad \mu_R(\pi) \geq \widetilde{G}(R, \pi) = [\pi]_1 V(x_1) + \cdots + [\pi]_n V(x_n) \, . \tag{41}$$

Optimal index strategies

Proposition 5

A optimal locally informed driver

1. *Selects the safe road if and only if*

$$\mu_R(\widehat{\pi}_t^\lambda) \leq V(x_S) \, . \tag{42}$$

2. *Sticks to the safe road, once he has selected it (a locally informed driver never switches from the safe road to the random road).*

The second assertion is rather intuitive since once a driver switches to a safe road, he does not update his information, so that there is no reason to shift back to the random road.

Proof.

1. This assertion is the major result on optimal strategies for bandit problems with independent arms and geometric discounting (see [7, 8, 16, 1]). Optimal strategies may be formulated as an "index strategy": pick up the arm with the higher index depending on the current state.

2. If a locally informed driver takes the safe road, then $\widehat{\pi}_{t+1}^\lambda = \widehat{\pi}_t^\lambda$ since there is no learning, hence no revision of the conditional law. Then, this driver sticks to the same choice since the state of the random road does not change. Thus, once a locally informed driver selects the safe road, he never switches back.

□

Notice that the optimal decision at period t only depends upon π_0 and upon the numbers N_t^1,…, N_t^n since we may write by Eq. (38):

$$\mu_R(\widehat{\pi}_t^\lambda) = \mu_R\left(\frac{\pi_0(dp_1 \cdots dp_n)p_1^{N_t^1-1} \cdots p_n^{N_t^n-1}}{\int_{S_{n-1}} \pi_0(dp_1 \cdots dp_n)p_1^{N_t^1-1} \cdots p_n^{N_t^n-1}}\right) \overset{\text{def}}{=} \overline{\mu}_R(\pi_0, N_t^+, N_t^-)\,. \tag{43}$$

5 A comparison of information regimes

We shall now detail properties of optimal strategies of the locally informed driver. It is a straightforward application of Proposition 5 that the optimal locally informed driver always selects the safe road if and only if $\mu_R(\pi_0) \leq V(x_S)$.

5.1 Locally versus globally informed driver

The road choice optimal behaviors of the locally and of the globally informed drivers are different but related in a way specified by Proposition below.

Proposition 6

1. *If an optimal locally informed driver always selects the safe road, he would also do so if he were globally informed.*
2. *If an optimal globally informed driver selects the random road from the first period up to a period t, he would do the same if he were locally informed and facing the same realizations of random travel times on the random road.*

Proof.

1. By Eq. (41) and by Eq. (29), if the locally informed driver selects the safe road at first period (and therefore at all periods), then so does the globally informed driver.
2. As a consequence, if the globally informed driver selects the random road at first period, then so does the locally informed driver. Then, their posteriors $\widehat{\pi}_t^\gamma$ and $\widehat{\pi}_t^\lambda$ coincide since they share the same prior and the same observations.

□

Note, however, that the optimal expected discounted rewards under probability $\mathbb{P}^{\overline{p}}$ cannot be ranked (while they may be ranked under probability $\mathbb{P}^{\pi_0}$ as in Eq. (45) below).

5.2 Comparison of optimal expected discounted rewards

We can now compare the optimal expected discounted rewards for the four information regimes envisaged.

Under the probability law $\mathbb{P}^{\overline{p}}$ which drives the realizations of random times on the random road, we have the following inequalities for optimal expected discounted rewards:

$$\mathbb{E}^{\overline{p}}[J(v^{v}(\cdot))] \geq \mathbb{E}^{\overline{p}}[J(v^{\phi}(\cdot))] \geq \max(\mathbb{E}^{\overline{p}}[J(v^{\gamma}(\cdot))], \mathbb{E}^{\overline{p}}[J(v^{\lambda}(\cdot))]) \,. \tag{44}$$

The first inequality has been shown in Paragraph 4.2. For the second, observe that the fully informed driver might use information $\mathcal{X}_t$, but it would not lead to a higher maximum since $\overline{p}$ is known. Thus $v^{\gamma}(\cdot)$ is a strategy admissible for the fully informed driver, hence the inequality $\mathbb{E}^{\overline{p}}[J(v^{\phi}(\cdot))] \geq \mathbb{E}^{\overline{p}}[J(v^{\gamma}(\cdot))]$. The inequality for $v^{\lambda}(\cdot)$ follows the same line of reasoning.

Note that there is no ranking for the last two information regimes. However, under the probability $\mathbb{P}^{\pi_0}$, we have that

$$\mathbb{E}^{\pi_0}[J(v^{\gamma}(\cdot))] \geq \mathbb{E}^{\pi_0}[J(v^{\lambda}(\cdot))] \,. \tag{45}$$

5.3 Risk aversion and locally informed drivers

The following Proposition corresponds to the general result of Proposition 1 and we simply sketch the proof.

Proposition 7
Consider two drivers with common belief π_0 and one more risk averse than the other. Assume that, at the beginning, the more risk averse driver selects the random road based on π_0. Then, so does the less risk averse driver and, as long as the more risk averse driver selects the random road, so does also the less risk averse driver.

Moreover, the mean time spent selecting the random road decreases with the degree of absolute risk aversion.

Proof. Assume that driver with utility function V^M is more risk averse than driver with utility function V^L. There exists a concave strictly increasing function φ such that $\varphi \circ V^L = V^M$. The state space is here $\mathbb{Z} = \mathcal{P}(S_{n-1})$ and the rewards are given by Eq. (20):

$$\Psi_S^{M,L} = V^{M,L}(x_S) \quad \text{and} \quad \Psi_R^{M,L}(\pi) = [\pi]_1 V^{M,L}(x_1) + \cdots + [\pi]_n V^{M,L}(x_n) \,, \quad \forall \pi \in \mathcal{P}(S_{n-1}) \,. \tag{46}$$

We have $\Psi_S^M = V^M(x_S) = \varphi(V^L(x_S)) = \varphi(\Psi_S^L)$. On the other hand, since φ is concave

$$\Psi_R^M(\pi) = [\pi]_1 \varphi(V^L(x_1)) + \cdots + [\pi]_n \varphi(V^L(x_n)) \leq \varphi([\pi]_1 V^L(x_1) + \cdots + [\pi]_n V^L(x_n)) = \varphi(\Psi_R^L(\pi)) \,.$$

The end of the proof follows with Proposition 10 in Appendix. □

These results are numerically illustrated in Section 6.

5.4 Risk aversion and certainty premium

In order to compare two optimal strategies corresponding to different information regimes, we use the concept of *certainty premium* [13, 9, 5].

Definition 8
To the driver with prior π_0 *and strategy* $v(\cdot)$, *we associate the* certainty premium $\Delta_{\pi_0,v(\cdot)}$ *implicitely defined by:*

$$V(x_S + \Delta_{\pi_0,v(\cdot)}) = (1-\rho)\mathbb{E}^{\pi_0}[\sum_{t=0}^{+\infty} \rho^t V(\Phi(v_t, X_{t+1}))] \,. \tag{47}$$

This is the additional travel time that he is willing to incur on the safe road to reach the level of discounted payoff that he gets with prior π_0 and strategy $v(\cdot)$.

Risk aversion and certainty premium are related as follows.

Proposition 9
Consider two drivers with common prior π_0 *and strategy* $v(\cdot)$. *If the driver with utility function* V^M *is more risk averse than the driver with utility function* V^L, *then* $\Delta^L_{\pi_0,v(\cdot)} \leq \Delta^M_{\pi_0,v(\cdot)}$.

Proof. Introduce the probability space $\mathbb{N} \times \Omega$ with probability $\mathbb{Q} \overset{\text{def}}{=} (1-\rho)\sum_{t=0}^{+\infty} \rho^t \delta_t \otimes \mathbb{P}^{\pi_0}$. Denoting $f(t,\omega) = \Phi(v_t(\omega), X_{t+1}(\omega))]$, we have $\mathbb{E}^{\pi_0}[\sum_{t=0}^{+\infty} \rho^t V(\Phi(v_t, X_{t+1}))] = \mathbb{E}_{\mathbb{Q}}(V(f))$. Thus, we are now in the classical framework to apply well known results on risk aversion. There exists a concave strictly increasing function φ such that $\varphi \circ V^L = V^M$, so that

$$\begin{aligned}
\varphi(V^L(x_S + \Delta^M_{\pi_0,v(\cdot)})) &= V^M(x_S + \Delta^M_{\pi_0,v(\cdot)}) \;\text{ since } \varphi \circ V^L = V^M \\
&= \mathbb{E}_{\mathbb{Q}}(V^M(f)) \;\text{ by definition of } \Delta^M_{\pi_0,v(\cdot)} \\
&= \mathbb{E}_{\mathbb{Q}}(\varphi(V^L(f))) \\
&\leq \varphi(\mathbb{E}_{\mathbb{Q}}(V^L(f))) \;\text{ by concavity of } \varphi \\
&= \varphi(V^L(x_S + \Delta^L_{\pi_0,v(\cdot)})) \,.
\end{aligned}$$

Since φ is strictly increasing and V is strictly decreasing, we conclude that $\Delta^M_{\pi_0,v(\cdot)} \geq \Delta^L_{\pi_0,v(\cdot)}$. □

The premium of the globally informed driver is Δ_γ such that

$$V(x_S + \Delta_\gamma) = (1-\rho) \sup_{v_t \preceq \mathcal{X}_t, t\geq 0} \mathbb{E}^{\pi_0}[\sum_{t=0}^{+\infty} \rho^t V(\Phi(v_t, X_{t+1}))] \,. \tag{48}$$

The premium of the locally informed driver is Δ_λ such that

$$V(x_S + \Delta_\gamma) = (1-\rho) \sup_{v_t \preceq \mathcal{Y}_t, t\geq 0} \mathbb{E}^{\pi_0}[\sum_{t=0}^{+\infty} \rho^t V(\Phi(v_t, X_{t+1}))] \,. \tag{49}$$

As a consequence of Eq. (45), together with the fact that V is decreasing, we have:

$$\Delta_\lambda \geq \Delta_\gamma \,. \tag{50}$$

Optimal locally informed drivers are eager to get a safe choice than optimal globally informed drivers. The difference $\Delta_\lambda - \Delta_\gamma$ measures what the driver is willing to pay in order to have access *ex post* to daily information on the travel times on both roads.

Combining ranking on information regimes and ranking with risk aversion, we obtain the following ranking of the premium:

$$\Delta_\lambda^M \geq \sup(\Delta_\lambda^L, \Delta_\gamma^M) \geq \inf(\Delta_\lambda^L, \Delta_\gamma^M) \geq \Delta_\gamma^L \, . \tag{51}$$

More precisely, the most risk averse driver is willing to pay the most for certainty if he is locally informed, and the least risk averse driver is willing to pay the least for certainty if he is globally informed.

6 Numerical illustration

For this numerical illustration, we restrict to the case $n = 2$. The random road R has a random travel time X_{t+1} realized at the end of period t which takes value in $\{x_-, x_+\}$ $(x_- < x_+)$. We suppose that x_- occurs with probability $\overline{p} \in]0,1[$, which is not known to the driver. We assume that

$$x_- < x_S < x_+ \, . \tag{52}$$

6.1 Equivalence with the one-armed Bernoulli problem

A one-armed Bernoulli problem is one in which the arm returns a Bernoulli random variable each time it is selected. The arm returns 1 with unknown probability $\overline{p}$ and 0 else. For any $\pi \in \mathcal{P}([0,1])$, let $\mu(\pi)$ denote the Gittins index for a Bernoulli arm. We shall now show how the Gittins index $\mu_R(\pi)$ of our random road may be easily expressed by means of μ.

Recall that the utility functions are defined up to a transformation $V \hookrightarrow aV + b$ $(a > 0)$, and let consider the following values

$$a = \frac{1}{V(x_-) - V(x_+)} \quad \text{and} \quad b = -\frac{V(x_+)}{V(x_-) - V(x_+)} \tag{53}$$

which are such that

$$aV(x_+) + b = 0 \quad \text{and} \quad aV(x_-) + b = 1 \, . \tag{54}$$

By Eq. (40), we have that $\mu(\pi) = a\mu_R(\pi) + b$, so that the safe road is selected if and only if

$$\mu(\pi) \leq \frac{V(x_S) - V(x_+)}{V(x_-) - V(x_+)} \, . \tag{55}$$

As expected, the right hand side of the above equation is increasing with risk aversion.

Such equivalence is possible only because the random variable takes no more than two values. Therefore, if $n \geq 2$ the risk aversion parameter adds a new dimension in the bandit problem.

6.2 Numerical computation of the Gittins index

This section is devoted to the numerical computation of the Gittins index μ_R for the random road.

By Eq. (41), and with obvious notations, we get that $\mu_R(\pi) \geq [\pi]_+ V(x_+) + [\pi]_- V(x_-)$. When π is a beta law, we have $[\beta(r,s)]_+ = r/(r+s)$ and $[\beta(r,s)]_- = s/(r+s)$, so that:

$$\mu_R(\beta(r,s)) \geq \frac{r}{r+s} V(x_+) + \frac{s}{r+s} V(x_-)\,. \tag{56}$$

The Gittins index defined by Eq. (1) may also be computed by the following dynamic programming scheme (see [2]). Let F_+ and F_- be the following "shift" operators on $\mathcal{P}([0,1])$:

$$F_+(\pi) \overset{\text{def}}{=} \frac{p\pi(dp)}{\int_0^1 p\pi(dp)}\,, \quad F_-(\pi) \overset{\text{def}}{=} \frac{(1-p)\pi(dp)}{\int_0^1 (1-p)\pi(dp)}\,. \tag{57}$$

The optimal utility $V_S(\pi, m)$, defined on $\mathcal{P}(S_1) \times \mathbb{R}_+$, for the problem on the safe road with retirement reward m satisfies

$$\begin{aligned} V_R(\pi, m) = \max\{m, \quad [\pi] \quad & V(x_n) + (1-[\pi])V(x_1) + \rho([\pi]V_R(F_+(\pi), m) \\ + \quad & (1-[\pi])V_R(F_1(\pi), m))\}\,. \end{aligned} \tag{58}$$

The index function μ_R is related to the optimal utility $V_S(\pi, m)$ by the relation (see [7, 2])

$$\frac{1}{(1-\rho)} \mu_R(\pi) = \min\{m \mid V_R(\pi, m) = m\} = \inf \arg\min_m [V_R(\pi, m) - m]\,. \tag{59}$$

It appears that the value function $V_R(\pi, m)$ satisfies a recursion when π is a beta law. Indeed, using the fact that $F_+(\beta(r,s)) = \beta(r+1, s)$, $F_1(\beta(r,s)) = \beta(r, s+1)$ and $[\beta(r,s)] = r/(r+s)$, we introduce $v_m(r,s) \overset{\text{def}}{=} V_R(\beta(r,s), m)$ and rewrite (58) as

$$v_m(r,s) = \max\{m, \frac{r}{r+s} V(x_n) + \frac{s}{r+s} V(x_1) + \rho[\frac{r}{r+s} v_m(r+1, s) + \frac{r}{r+s} v_m(r, s+1)]\}\,. \tag{60}$$

Thus, the computation of $v_m(r,s)$ requires the values of $v_m(r+1,s)$ and $v_m(r,s+1)$. As a consequence, one is led to compute $v_m(r', s')$ on the grid $\{(r+n, s+m), n \in \mathbb{N}, m \in \mathbb{N}\}$.

First, let us explain how $v_m(r,s)$ is evaluated on the grid $\mathbb{N}_k \overset{\text{def}}{=} \{0, 1, ..., k\}^2$. Equation (60) is a fixed point equation which can be writen as $v_m = \mathcal{L}(v_m)$, where $\mathcal{L}$ is a strict contraction ($\rho < 1$). The numerical scheme used is a value iteration algorithm: given v_m^0 we compute $v_m^i = \mathcal{L}(v_m^{i-1})$. But we also have to localize the problem since it is not possible to iterate $\mathcal{L}$ for a function defined on the whole $\mathbb{N} \times \mathbb{N}$. Equation (60) shows that computing v_m^i on $\mathbb{N}_k$ can be done if we have already computed v_m^{i-1} on $\mathbb{N}_{k+1}$. Thus, in order to perform $N = 100$ iterations of the value iteration algorithm, we start the algorithm with $v_m \overset{\text{def}}{=} m$ on $\mathbb{N}_{50+N}$ and the fixed point is approximated by v_m^N computed on $\mathbb{N}_{50}$. Note that the value iteration algorithm in that case computes a finite horizon ($N = 100$) approximation of a stopping time problem. The v_m functions are computed for a discretized set of m values and then μ_R is computed using equation Eq. (59).

Second, we may refine the grid on which the v_m functions are evaluated by repeating the same scheme on the grid $(1/2, 1/2) + \mathbb{N}_k \overset{\text{def}}{=} \{1/2, 3/2, ..., k + 1/2\}^2$. The refinement may go on.

For numerical experiments we have used $\rho = 1/1.08$ and the following CARA utility function

$$V_\theta(x) = \frac{(1 - e^{\theta x})}{\theta} \tag{61}$$

with $x^- = 10/60$, $x_s = 20/60$ and $x^+ = 22/60$. The parameter θ is the *Arrow-Pratt degree of absolute risk aversion* $-V_\theta''/V_\theta'$.

6.3 Numerical results

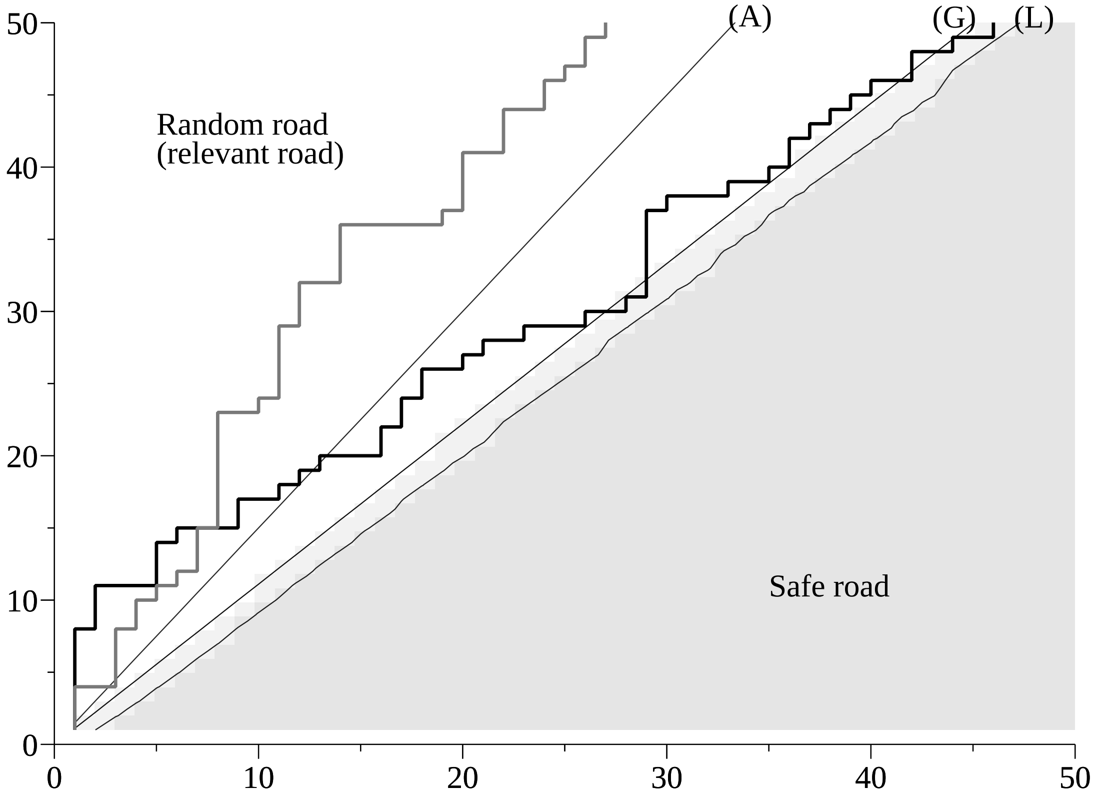

Figure 1: Trajectories of posterior laws when the relevant road is the random road

Figures 1 and 2 show some typical trajectories of $\widehat{\pi}_t$ (posterior laws) for a starting value $\widehat{\pi}_0 = \beta(1,1)$ in a two dimensional space where the horizontal axis correspond to the number of bad states (high travel time) and the vertical axis correspond to the number of good states.

The points below the (L) curve define an area (the gray zone) for which $\mu_S \geq \mu_R$. Thus, as long as the current point of a trajectory remains above the (L) curve, the optimal strategy

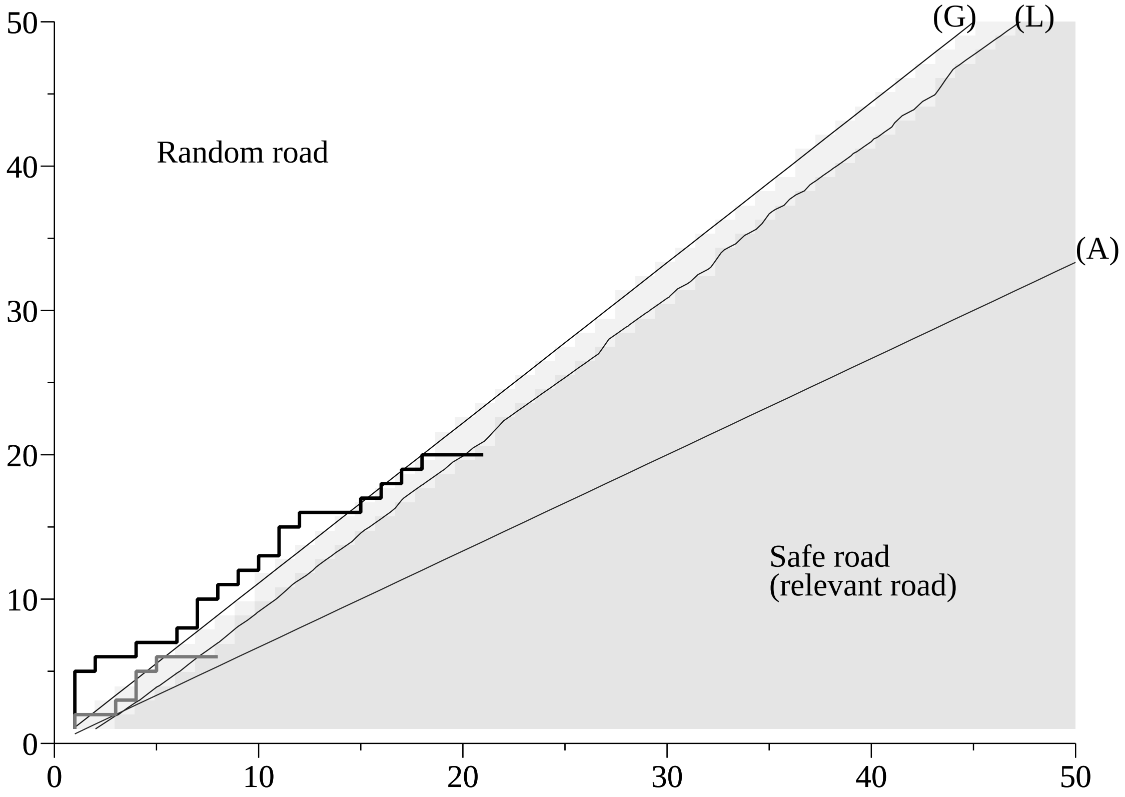

Figure 2: Trajectories of posterior laws when the relevant road is the safe road

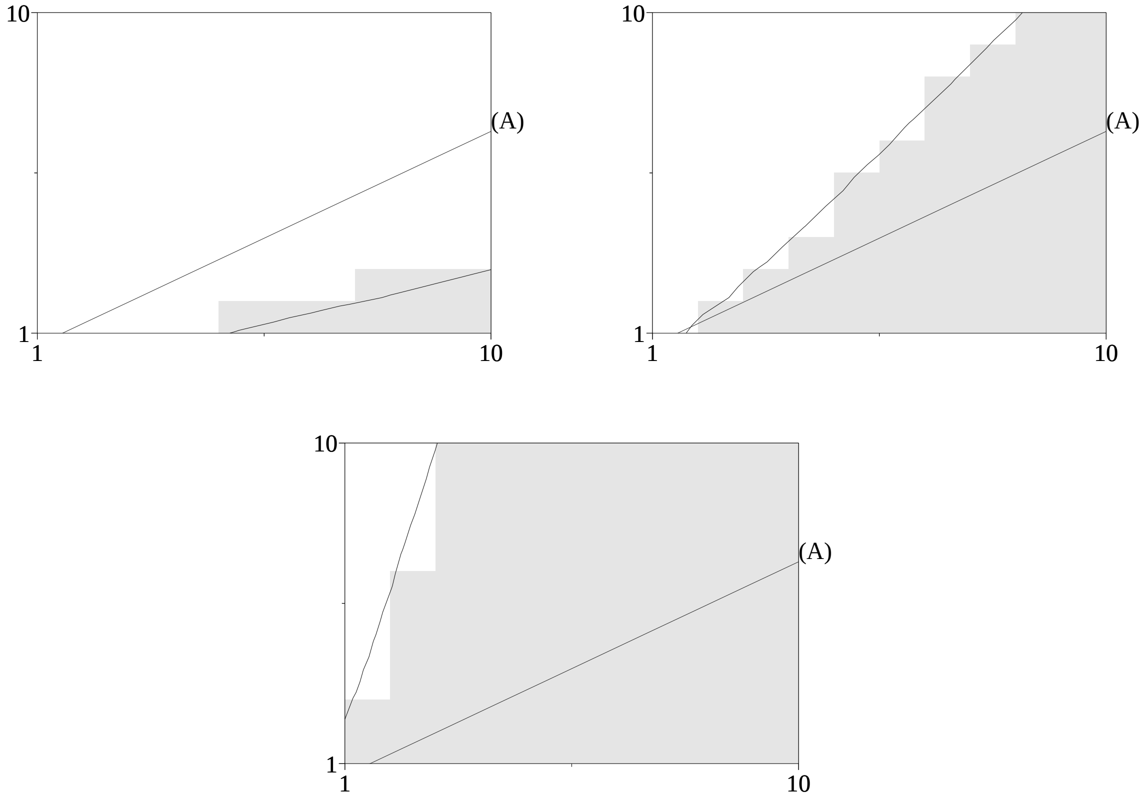

Figure 3: Gittins rules for increasing values of θ (7, 27, 53)

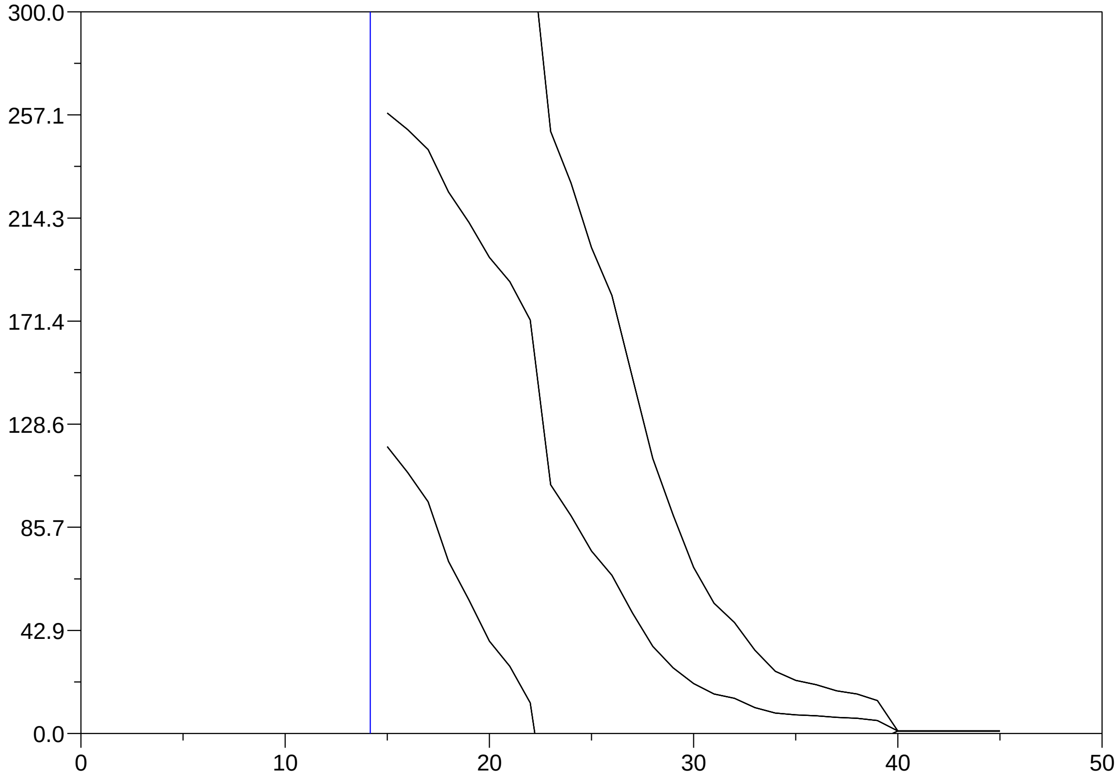

Figure 4: Mean time spent on the random road (± one standart deviation) as function of θ

is to stay on the random road R. When a trajectory hits the (L) curve (if it does), it is then optimal to switch to the safe road S; from then on, the driver will optimally stay on the safe road and his posterior law remains constant.

The points below the (G) curve (the union of the light gray and gray zones) describe the area for which $\mu_S(\beta(r,s)) \geq \widetilde{G}(R, \beta(r,s))$ (see Eq. (41)). This stopping rule is easy to compute and provides a reasonably good approximation of the Gittins index rule, for the parameter values considered here. Note that (G) and (L) are asymptotically parallel lines.

As the number of iteration tends to infinity, the trajectories of the posterior laws for the globally informed driver asymptotically converge to the (A) curve (see Eq. (35)). In Figure 1, the (A) curve is above the (L) curve. Thus the relevant road is the random one and each trajectory of the locally informed driver which hits the (L) curve leads to a stationary choice (S) which is not relevant. In Figure 2, the (A) curve is below the (L) curve. The relevant road is the safe road and, in this case, the trajectories hit the (L) curve almost surely (and therefore the asymptotic choice coincide a.s with the relevant choice).

Figure 3 shows that the safe region (gray zone) gets larger for increasing values of risk aversion, θ, highlighting numerically Proposition 1. We represent in the same space as for Figure 1 and 2 the (L) curve and the gray zone on the grid $\mathbb{N}_{10}$. For $\theta = 7$ (top left sub-figure) the relevant road is the random road and the mean time on the random road is infinite. For the two other sub-figures, the relevant road is the safe road: for $\theta = 27$ the mean time is positive while, in the second case $\theta = 53$, the mean time is null.

This discussion is further illustrated by Figure 4 where we present the mean time on the random road (plus or minus one standard deviation) as a function of the risk aversion parameter θ. The vertical asymptote is at $\theta^\star = 14.7$. It corresponds to the individual with risk aversion $\theta^\star$ which makes him indifferent between the safe and the random road given the probability $\overline{p}$. A driver with $\theta < \theta^\star$ always sticks to the random road while the driver with $\theta \in]\theta^\star, 40[$ starts to use the random road but, sooner or later, ends up on the safe road. Individuals with risk aversion larger than 40 never try the random road. Note that this critical value, denoted $\overline{\theta}$, of risk aversion is such that $\overline{\theta} = \inf\{\theta \mid \mu_R(1,1) < \mu_S\}$ where $(1,1)$ stands for the uniform law. Numerically, the mean time spent on the random road is obtained by Monte Carlo simulations and, since each simulated trajectory is stopped on the boundary of the domain on which the Gittins index is computed, the mean time is underestimated.

7 Conclusion

We have studied the choice between a safe and a random alternative under four information regimes: the most difficult one (locally informed users) is when users need to select an alternative in order to get information about it. The scope of this paper was to study the impact of individual risk aversion in such situations. In Section 6, we have shown how the one-armed bandit *with* risk aversion may be reduced to one-armed bandit *without* risk aversion if the random arm can be in two states only. With more than two states or more than a random arm, risk aversion modifies the solution of the armed bandit problem in a non-trivial manner. In particular, we show that a more risk averse user has more chance to select at any time the random alternative than a less risk averse user. We also show that a more risk averse individual is willing to pay more in order to have access to a safe alternative. However, we argue below that this does not mean that more risk averse users are all willing to experiment more in the one-armed bandit problem.

The way individuals wish to reduce uncertainty as a function of their risk aversion is not trivial. Casual intuition (and our initial intuition) suggests that more risk averse users may be willing to invest more to reduce uncertainty. We have shown that more risk averse individuals actively tend to invest more effort to reduce uncertainty, in a subtle manner described below. In our numerical illustration, individuals who are almost risk neutral, or have a risk aversion less that a threshold $\theta^\star$ select the random road and stick to it. They reduce uncertainty over time, but this information is less and less relevant in the sense that their initial and final choice coincide (relevant road); note that, over time, this initial choice has a smaller and smaller probability to be overturned. Conversely, the individuals with a high degree of risk aversion ($\theta > \overline{\theta}$) do not wish to know the characteristic of the random road, and select right away the safe choice (and therefore never learn anything). The individuals who are more risk averse than $\theta^\star$, but less risk averse than $\overline{\theta}$, initially select the random road to get information, and after a certain number of iterations, which decreases with their degree of risk aversion, shift to the safe (relevant) road (see Figure 4). What matters are the individuals who would be indifferent between the safe and the random choice, if they knew about $\overline{p}$. Those users are the ones who need to learn the most before they can make their final choice.

The situation could be better explained in an extended model, where it would be costly to process the information (e.g. to hire a consulting firm to make predictions). The two processes – acquisition and processing (Bayesian update) of information – are described within a single process in the classical bandit problem. Yet, a user may acquire information but he may not be willing to put effort to process it. We conjecture that the individuals who are at the vicinity of $\overline{\theta}$ are willing to invest the most in order to process the information they acquire (although this is debatable in the sense that in this case, the safe and the random roads lead to similar rewards).

Road choice and drivers information system has generated a very large literature in the transportation field, and to a less extent in economics. Those studies resort to simulation measuring the impacts of information on congestion and do not capture the preferences of individuals with respect to risk. Here we presented a very different view since we analyze the impact of information on rewards, including the benefits from reducing uncertainty (see also [4] which considers decisions of fully informed drivers with CARA or CRRA preferences). We believe that our approach with learning will be useful to study the economic benefit of information. However, these problems request that the bandit model should be extended to situation where the benefit of one individual depends on the choice of others (this interaction corresponds to congestion).

Finally, our model can be understood also in the context of discrete choice models. However, learning and uncertainty have not been described in this literature. The bandit approach suggests a discrete choice model with rational learning. We believe that this approach could be used as a basis to develop structural dynamic discrete choice models, for repetitive choices where leaning takes place (going to the restaurant, going to a shopping mall, selecting a mileage plus company, etc.). The theory and the econometric properties of those models, which combine the modeler's uncertainty (the error term in the utility function) and the individual's uncertainty (inherent to the classical bandit problem) are still to be envisaged.

A Lemmas and proofs

A.1 Comparison of bandit problems

Consider one decision-maker (M) which faces a one-armed bandit problem. The safe arm S returns, when selected, a deterministic fixed reward $\Psi_S^M \in \mathbb{R}$. The random arm R has state space $\mathbb{Z}$ and reward $\Psi_R^M : \mathbb{Z} \to \mathbb{R}$. A transition kernel is given on $\mathbb{Z}$ and, when the random arm R is selected at period t, its state moves from z_t towards z_{t+1} according to this transition kernel. Defining $\Psi^M : \{S, R\} \times \mathbb{Z} \to \mathbb{R}$ by $\Psi^M(S, z) \overset{\text{def}}{=} \Psi_S^M$ and $\Psi^M(R, z) \overset{\text{def}}{=} \Psi_R^M(z)$, the decision-maker (M) has to solve

$$\sup_{v(\cdot)} \mathbb{E}[\sum_{t=0}^{\infty} \rho^t \Psi^M(v_t, z_t)] \tag{62}$$

where $\rho \in [0, 1[$ is the discount rate and the law of z_0 is given (which determines, with the transition kernel, the probability $\mathbb{P}$ corresponding to the mathematical expectation $\mathbb{E}$). Here, the strategy $v(\cdot)$ is such that v_t may depend upon $z_0, \ldots, z_t$ assumed to be observed.

Now, consider another decision-maker (L) which faces the same one-armed bandit, except for the rewards. With obvious notations, the rewards are $\Psi_S^L \in \mathbb{R}$ and $\Psi_R^L : \mathbb{Z} \to \mathbb{R}$.

We compare the optimal strategies of these two decision-makers (M) and (L) (More and Less). Our main result concerning the impact of risk aversion on optimal strategies in a one-armed bandit problem is the following.

Proposition 10
Assume there exists a concave increasing function $\varphi : \mathbb{R} \to \mathbb{R}$ such that

$$\Psi_S^M \geq \varphi(\Psi_S^L) \quad \text{and} \quad \Psi_R^M(z) \leq \varphi(\Psi_R^L(z)) \quad \forall z \in \mathbb{Z}\,. \tag{63}$$

Then, each time the agent with rewards (Ψ_R^M, Ψ_S^M) selects the random arm, so does the agent with rewards (Ψ_R^L, Ψ_S^L) when he is in the same state.

As a straightforward corollary, each time the agent with rewards (Ψ_R^L, Ψ_S^L) selects the safe arm, so does the agent with rewards (Ψ_R^M, Ψ_S^M) when he is in the same state. However, we are unable to identify assumptions ensuring that each time the agent with rewards (Ψ_R^M, Ψ_S^M) selects the *safe* arm, so does the agent with rewards (Ψ_R^L, Ψ_S^L) when he is in the same state.

Proof. By definition Eq. (1), we have

$$\begin{cases} \mu_S^{M,L}(z) &= \sup_{\tau>0} \dfrac{\mathbb{E}[\sum_{t=0}^{\tau-1} \rho^t \Psi_S^{M,L} \mid z_0 = z]}{\mathbb{E}[\sum_{t=0}^{\tau-1} \rho^t \mid z_0 = z]} = \Psi_S^{M,L} \\ \mu_R^{M,L}(z) &= \sup_{\tau>0} \dfrac{\mathbb{E}[\sum_{t=0}^{\tau-1} \rho^t \Psi_R^{M,L}(z_t) \mid z_0 = z]}{\mathbb{E}[\sum_{t=0}^{\tau-1} \rho^t \mid z_0 = z]}\,. \end{cases} \tag{64}$$

Let $\tau > 0$ be a fixed stopping time. We introduce the random variable $Y = \sum_{t=0}^{\tau-1} \rho^t > 0$ and a new probability $\widetilde{\mathbb{P}}$ such that $\widetilde{\mathbb{E}}(X) = \frac{\mathbb{E}(XY|z_0=z)}{\mathbb{E}(Y|z_0=z)}$. We have

$$
\begin{aligned}
\frac{\mathbb{E}[\sum_{t=0}^{\tau-1} \rho^t \Psi_R^M(z_t) \mid z_0 = z]}{\mathbb{E}[\sum_{t=0}^{\tau-1} \rho^t \mid z_0 = z]} &\leq \frac{\mathbb{E}[\sum_{t=0}^{\tau-1} \rho^t \varphi(\Psi_R^L(z_t)) \mid z_0 = z]}{\mathbb{E}[\sum_{t=0}^{\tau-1} \rho^t \mid z_0 = z]} \text{ since } \Psi_R^M \leq \varphi \circ \Psi_R^L \\
&= \widetilde{\mathbb{E}}[\sum_{t=0}^{\tau-1} \rho^t \frac{\varphi(\Psi_R^L(z_t))}{\sum_{s=0}^{\tau-1} \rho^s}] \text{ by definition of } \widetilde{\mathbb{E}} \\
&\leq \widetilde{\mathbb{E}}[\varphi(\sum_{t=0}^{\tau-1} \rho^t \frac{\Psi_R^L(z_t)}{\sum_{s=0}^{\tau-1} \rho^s})] \text{ since } \varphi \text{ is concave} \\
&\leq \varphi(\widetilde{\mathbb{E}}[\sum_{t=0}^{\tau-1} \rho^t \frac{\Psi_R^L(z_t)}{\sum_{s=0}^{\tau-1} \rho^s}]) \\
&\qquad \text{by Jensen inequality, since } \varphi \text{ is concave} \\
&= \varphi(\frac{\mathbb{E}[\sum_{t=0}^{\tau-1} \rho^t \Psi_R^L(z_t) \mid z_0 = z]}{\mathbb{E}[\sum_{t=0}^{\tau-1} \rho^t \mid z_0 = z]}) \text{ by definition of } \widetilde{\mathbb{E}}.
\end{aligned}
$$

Thus, $\mu_R^M(z) \leq \varphi(\mu_R^L(z))$ since

$$
\begin{aligned}
\mu_R^M(z) &= \sup_{\tau>0} \frac{\mathbb{E}[\sum_{t=0}^{\tau-1} \rho^t \Psi_R^M(z_t) \mid z_0 = z]}{\mathbb{E}[\sum_{t=0}^{\tau-1} \rho^t \mid z_0 = z]} \\
&\leq \sup_{\tau>0} \varphi(\frac{\mathbb{E}[\sum_{t=0}^{\tau-1} \rho^t \Psi_R^L(z_t) \mid z_0 = z]}{\mathbb{E}[\sum_{t=0}^{\tau-1} \rho^t \mid z_0 = z]}) \\
&\leq \varphi(\sup_{\tau>0} \frac{\mathbb{E}[\sum_{t=0}^{\tau-1} \rho^t \Psi_R^L(z_t) \mid z_0 = z]}{\mathbb{E}[\sum_{t=0}^{\tau-1} \rho^t \mid z_0 = z]}) \text{ since } \varphi \text{ is increasing} \\
&= \varphi(\mu_R^L(z)) .
\end{aligned}
$$

Now, we have by assumption $\mu_S^M(z) = \Psi_S^M \geq \varphi(\Psi_S^L) = \varphi(\mu_S^L(z)$, so that

$$
\begin{aligned}
\mu_R^M(z) \geq \mu_S^M(z) &\Rightarrow \mu_R^M(z) \geq \varphi(\mu_S^L(z)) \text{ since } \mu_S^M(z) \geq \varphi(\mu_S^L(z)) \\
&\Rightarrow \varphi(\mu_R^L(z)) \geq \varphi(\mu_S^L(z)) \text{ since } \varphi(\mu_R^L(z)) \geq \mu_R^M(z) \\
&\Rightarrow \mu_R^L(z) \geq \mu_S^L(z) \text{ since } \varphi \text{ is increasing.}
\end{aligned}
$$

As a consequence, when the agent with rewards (Ψ_R^M, Ψ_S^M) selects the random arm, so does the agent with rewards (Ψ_R^L, Ψ_S^L) when he is in the same state. This ends the proof. □

A.2 Technical lemma

Lemma 11

Let $(\mathcal{Z}_t)_{t\in\mathbb{N}}$ be a family of subfields of the σ-field $\mathcal{F} = 2^\Omega$. Let $v(\cdot) = (v_0, v_1, \ldots)$ be a sequence of decisions such that $v_t \preceq \mathcal{Z}_t$ (meaning that v_t is $\mathcal{Z}_t$-measurable). Then

$$\mathbb{E}^{\pi_0}[J(v(\cdot))] = \mathbb{E}^{\pi_0}[\sum_{t=0}^{+\infty} \rho^t \widetilde{G}(v_t, \widehat{\pi}_t)] \tag{65}$$

where $\widetilde{G}$ is given by Eq. (20) and $\widehat{\pi}_t \in \mathcal{P}(S_{n-1})$ is any probability law on S_{n-1} such that

$$\forall i = 1, \ldots, n\,, \quad [\widehat{\pi}_t]_i = \mathbb{P}^{\pi_0}(X_{t+1} = x_i \mid \mathcal{Z}_t)\,. \tag{66}$$

Proof. Let $\widehat{\pi}_t$ be such that Eq. (66) holds. We have then for $v \in \{R, S\}$:

$$\begin{aligned}
\mathbb{E}^{\pi_0}[G(v, X_{t+1}) \mid \mathcal{Z}_t] &= \mathbb{P}^{\pi_0}(X_{t+1} = x_1 \mid \mathcal{Z}_t)G(v, x_1) + \cdots + \mathbb{P}^{\pi_0}(X_{t+1} = x_n \mid \mathcal{Z}_t)G(v, x_n) \\
&= [\widehat{\pi}_t]_1 G(v, x_1) + \cdots + [\widehat{\pi}_t]_n G(v, x_n) \quad \text{by Eq. (66)} \\
&= \mathbb{E}^{\widehat{\pi}_t}[\widetilde{G}(v_t, \widehat{\pi}_t)] \quad \text{by Eq. (20).}
\end{aligned}$$

Thus,

$$\begin{aligned}
\mathbb{E}^{\pi_0}[J(v(\cdot))] &= \mathbb{E}^{\pi_0}[\sum_{t=0}^{+\infty} \rho^t G(v_t, X_{t+1})] \quad \text{by Eq. (11)} \\
&= \sum_{t=0}^{+\infty} \rho^t \mathbb{E}^{\pi_0}[G(v_t, X_{t+1})] \\
&= \sum_{t=0}^{+\infty} \rho^t \mathbb{E}^{\pi_0}[\mathbb{E}^{\pi_0}[G(v_t, X_{t+1}) \mid \mathcal{Z}_t]] \\
&= \sum_{t=0}^{+\infty} \rho^t \mathbb{E}^{\pi_0}[\mathbb{E}^{\pi_0}[G(v, X_{t+1}) \mid \mathcal{Z}_t]_{|v=v_t}] \quad \text{since} \quad v_t \preceq \mathcal{Z}_t \\
&= \sum_{t=0}^{+\infty} \rho^t \mathbb{E}^{\pi_0}[\widetilde{G}(v_t, \widehat{\pi}_t)] \\
&= \mathbb{E}^{\pi_0}[\sum_{t=0}^{+\infty} \rho^t \widetilde{G}(v_t, \widehat{\pi}_t)]\,.
\end{aligned}$$

□

References

[1] D. A. Berry and B. Fristedt. *Bandit problems: sequential allocation of experiments.* Chapman and Hall, 1985.

[2] D. P. Bertsekas. *Dynamic Programming and Optimal Control*, volume 1 and 2. Athena Scientific, Belmont, Massachusets, second edition, 2000.

[3] D. P. Bertsekas and S. E. Shreve. *Stochastic Optimal Control: The Discrete-Time Case.* Athena Scientific, Belmont, Massachusets, 1996.

[4] A. de Palma and N. Picard. Congestion on risky routes with risk averse drivers. *Working paper* THEMA, Université of Cergy-Pontoise, France, 2005.

[5] P. Diamond and M. Rothschild, Editors. *Uncertainty in Economics.* Academic Press, Orlando, 1978.

[6] D. Ellsberg. Risk, ambiguity, and the Savage axioms. *Quartely Journal of Economics*, 75:643–669, 1961.

[7] J. C. Gittins. Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society. Series B*, 41(2):148–177, 1979.

[8] J. C. Gittins. *Multi-armed Bandit Allocation Indices.* Wiley, New York, 1989.

[9] C. Gollier. *The Economics of Risk and Time.* MIT Press, Cambridge, 2001.

[10] D. Kahneman and A. Tversky. Prospect theory: an analysis of decision under risk. *Econometrica*, 47:263–291, 1979.

[11] D. McFadden. Economic choices. *The American Economic Review*, 91(3):351–378, June 2001.

[12] J. Pitman. *Probability.* Springer-Verlag, New-York, 1993.

[13] J. W. Pratt. Risk aversion in the small and in the large. *Econometrica*, 32(1-2):61–75, 1964.

[14] S. P. Anderson, A. de Palma and J.-F. Thisse. *Discrete Choice Theory of Product Differentiation.* MIT Press, Cambridge, 1992.

[15] J. von Neuman and O. Morgenstern. *Theory of Games and Economic Behaviour.* Princeton University Press, Princeton, 1947. 2nd edition.

[16] P. Whittle. *Optimization over Time: Dynamic Programming and Stochastic Control.* John Wiley & Sons, New York, 1982.