Risk aversion in expected intertemporal discounted utilities bandit problems

Jean-Philippe Chancelier†, Michel De Lara†, André de Palma‡

†CERMICS, École des ponts, Paris Tech, 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 Institut universitaire de France

June 11, 2007

1 Introduction

Consider a one-armed bandit problem as follows. The agent selects between a certain and a random arm, which yields a gain G_t at time t. Most of the literature focuses on maximizing $\mathbb{E}[\sum_{t=0}^{\infty} \rho^t G_t]$, that is intertemporal discounted expected reward. Now, assume that the agent evaluates the gain by means of a utility function V, giving a new reward $V(G_t)$ measured in utility. Though disputable and discussed in the economic literature [7], it is widely accepted to maximize $\mathbb{E}[\sum_{t=0}^{\infty} \rho^t V(G_t)]$, that is *expected intertemporal discounted utilities* (of rewards). We study how the agent risk aversion – measured by the degree of concavity of the utility function V (related to Arrow-Pratt absolute risk aversion) – affects the sequence of arms selection. We show that the more the utility function is convave, the more the agent selects the certain arm. This result is the consequence of our main result as to how optimal decisions vary when the rewards[1] vary.

To the best of our knowledge, the question we raise has not been treated in this form in the literature. This does not mean that the question of incorporating risk in bandit problems has not been examined, particularly in the literature of operations research in the so called *risk sensitive* approach. In this latter case, you first evaluate intertemporal (non discounted) rewards $\sum_{t=0}^{T} G_t$ up to horizon T, second measure it in (exponential) utility[2], third take the expectation, fourth take a certainty equivalent giving $-\frac{1}{\theta} \log \mathbb{E}\big[\exp(-\theta \sum_{t=0}^{T-1} G_t)\big]$. Thus, after normalizing with time T, the criterion to be maximized is a limit when T goes to infinity of $-\frac{1}{\theta T} \log \mathbb{E}\big[\exp(-\theta \sum_{t=0}^{T-1} G_t)\big]$, where $\theta > 0$ is a measure of risk aversion. In a

[1] This is another point of view than the one in [8], where monotonicity properties of the optimal decisions with respect to the probability distribution of the state process have been studied.

[2] For the specific CARA utility function $V_\theta(x) = -\frac{e^{-\theta x}}{\theta}$.

sense, such risk sensitive agents maximize *expected utility of intertemporal rewards* while we consider *expected intertemporal discounted utilities* of rewards $\mathbb{E}[\sum_{t=0}^{\infty} \rho^t V(G_t)]$ maximization (where V may be *any* utility function). A recent work on risk sensitive criteria and bandit problems is [3].

Consider the individual search for the best durable items. Assume that the goods searched are homogeneous but differ only along one characteristic distributed over items which can be: the price (search for a durable good), the wage (job search) or the location (search for residence). Each search involves a fixed cost and the distribution of characteristics is either known or unknown. In the latter case, after each costly examination of an item, the individual revise his posterior about the distribution of characteristic and either may decide to acquire this item (or an item already examined) or may decide to continue the search. The optimal stopping rule (with or without known distributions) is standard and has been studied for example by Rothschild [10]. Initially, individual were assumed to be risk neutral. Later on, several theoretical and empirical articles have considered also risk adverse decision makers: in this case, the optimal stopping rule depends on the level of risk aversion (see, for example the survey of Wolpin [12] for the study of the optimal dynamic fertility models).

Alternatively, the individual is searching for a non durable good (newspaper, restaurant or route from home to work) that he will consume/use repetitively. One characteristics of this good is stochastic, and can be sampled only when this good is consumed/used: the quality of information contained in a newspaper may vary from day to day, the quality of a chef is typically not constant (at least in good enough restaurant!), and the travel time on a route may also vary from day to day. Such goods are experience goods since the information about their unknown characteristic (quality of the news, style of the chef and the travel time, respectively) can only be updated after consumption. In this case, an individual who consumes the good acquires two bundled payoffs: the (stochastic) net surplus from this good and a realization of the unknown characteristic. This durable good problem differs from the non-durable good problem (discusses previously), since in the former case information acquisition has a constant cost, while in the latter it has an endogenous cost given that examination means consumption. The solution of this search model, with repetitive consumption has been worked out in clinical trials and in economics; it is known as the armed bandit problem (see [5, 6, 1]).

We concentrate our attention on non-durable good models. The literature assumed that the consumer are risk neutral, so that their objective is to maximize the expected outcome. For example, this means that in clinical trials the modeler searches for the drug with the minimal expected number of adverse effects, while in economics, the individual select a good which maximizes its average quality. However, uncertainty and risk aversion are inherent involved in those problems, so that risk neutrality may be restrictive. Strangely enough, to the best of our knowledge, no literature has been devoted to the consumption of non-durable leaning goods (repetitive choices) when risk averse individuals face uncertainty. The question we wish to solve is: what are the consequences of risk aversion in an armed bandit? Note that, on a related literature, the authors have recently introduced the bandit problem in the transportation science literature. In [2], they explore the day to day route choice problem

for a single individual facing the choice between a risky route (with certain travel time) and a safe route (with constant travel time).

We consider a simplified situation where the individual faces two choices, a safe choice with known characteristic and a risky choice, with unknown characteristics (given by some prior before any consumption has been made). This situation is particular since the individual learns nothing while acquiring the safe good. As a consequence, once he decides to select the safe good, he will continue to select it forever. So the individual may always select the safe good, or always select the risky good. A third solution is that he starts consuming the risky good and ends up selecting the safe good. Our purpose is to introduce risk aversion in this one armed bandit problem and to study how risk aversion modifies consumer behavior.

Consider for example an individual after high school who should decide to continue studying or not. The choice to continue to study or to go to the job market occurs at the end of each year. The student faces two sources of uncertainty: the probability of success and the (personal) benefits from studying one more year. Some students may have a high prior probability to fail, be very risk averse and go directly to the job market after high school. However, an equally capable student but less risk averse may stay and complete the program. Finally, another equally capable students, but with an intermediate level of risk aversion may start the university to see how well he performs and then decide to drop after a few years, when he finds out that his performance is not that great. The decision to study or not depends on a) his perceived probability of success (which gets from year to year more accurate) and on b) his level of risk aversion. When the job market is safe[3] (full employment and known wages), this is a one armed bandit problem with risk averse decision makers that we study here with a parameterized level of risk aversion.

In the next two sections, we compare the optimal strategies of two individuals facing a armed bandid problem, when they differ in their level of risk aversion. The main result of the paper is provided in Theorem 1 (Section 2), which is then applied in Proposition 2 (Section 3).

2 Comparison of rewards and strategies in a one-armed bandit problem

Our presentation of bandit problems is quite sketchy, and we send the reader to specialized references such as [5, 6, 11, 1].

Consider one decision-maker (M) which faces a one-armed bandit problem. The *certain arm* C returns, when selected, a deterministic fixed reward $\Psi_C^M \in \mathbb{R}$. The *random arm* R has *state space* $\mathbb{S}$ and reward $\Psi_R^M : \mathbb{S} \to \mathbb{R}$. A transition kernel is given on $\mathbb{S}$ and, when the random arm is selected at period t, its state moves from s_t towards s_{t+1} according to this transition kernel (that we do not need to specify in what follows). Defining $\Psi^M : \{C, R\} \times \mathbb{S} \to \mathbb{R}$ by $\Psi^M(C, s) := \Psi_C^M$ and $\Psi^M(R, s) := \Psi_R^M(s)$, the decision-maker ($M$) has

[3] If the job market is risky, transitions from the job market to the education system are also rational: this is a two armed bandit, not considered here.

to solve $\sup_{v(\cdot)} \mathbb{E}[\sum_{t=0}^{\infty} \rho^t \Psi^M(v_t, s_t)]$, where $\rho \in]0,1[$ is the discount rate and the law of s_0 is given. Here, the strategy $v(\cdot)$ is such that v_t may depend upon $s_0, \ldots, s_t$ 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 $\Psi_C^L \in \mathbb{R}$ and $\Psi_R^L : \mathbb{S} \to \mathbb{R}$. We compare the optimal strategies of these two decision-makers (M) and (L) (More and Less).

Theorem 1 *Assume that there exists a concave increasing function $\varphi : \mathbb{R} \to \mathbb{R}$ such that*

$$\Psi_C^M \geq \varphi(\Psi_C^L) \quad \textit{and} \quad \Psi_R^M(s) \leq \varphi\big(\Psi_R^L(s)\big) \quad \forall s \in \mathbb{S}\,. \tag{1}$$

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

As a straightforward corollary, each time the agent with rewards (Ψ_R^L, Ψ_C^L) selects the certain arm, so does the agent with rewards (Ψ_R^M, Ψ_C^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_C^M) selects the *certain* arm, so does the agent with rewards (Ψ_R^L, Ψ_C^L) when he is in the same state.

Proof. The Gittins indexes are the following supremum over stopping times $\tau > 0$ (see [5]):

$$\left\{ \begin{array}{rcl} \mu_C^{M,L}(s) & := & \displaystyle\sup_{\tau>0} \left(\frac{\mathbb{E}[\sum_{t=0}^{\tau-1} \rho^t \Psi_C^{M,L} \mid s_0 = s]}{\mathbb{E}[\sum_{t=0}^{\tau-1} \rho^t \mid s_0 = s]} \right) = \Psi_C^{M,L} \\ \\ \mu_R^{M,L}(s) & := & \displaystyle\sup_{\tau>0} \left(\frac{\mathbb{E}[\sum_{t=0}^{\tau-1} \rho^t \Psi_R^{M,L}(s_t) \mid s_0 = s]}{\mathbb{E}[\sum_{t=0}^{\tau-1} \rho^t \mid s_0 = s]} \right). \end{array} \right. \tag{2}$$

Let $\tau > 0$ be a fixed stopping time. We introduce the random variable $Y = \displaystyle\sum_{t=0}^{\tau-1} \rho^t > 0$ and a new

probability $\widetilde{\mathbb{P}}$ such that $\widetilde{\mathbb{E}}(X) = \frac{\mathbb{E}(XY|s_0=s)}{\mathbb{E}(Y|s_0=s)}$. We have

$$\begin{aligned}
\frac{\mathbb{E}[\sum_{t=0}^{\tau-1}\rho^t\Psi_R^M(s_t) \mid s_0=s]}{\mathbb{E}[\sum_{t=0}^{\tau-1}\rho^t \mid s_0=s]} &\leq \frac{\mathbb{E}[\sum_{t=0}^{\tau-1}\rho^t\varphi(\Psi_R^L(s_t)) \mid s_0=s]}{\mathbb{E}[\sum_{t=0}^{\tau-1}\rho^t \mid s_0=s]} \text{ since } \Psi_R^M \leq \varphi\circ\Psi_R^L \\
&= \widetilde{\mathbb{E}}[\sum_{t=0}^{\tau-1}\rho^t\frac{\varphi(\Psi_R^L(s_t))}{\sum_{s=0}^{\tau-1}\rho^s}] \text{ by definition of } \widetilde{\mathbb{E}} \\
&\leq \widetilde{\mathbb{E}}[\varphi(\sum_{t=0}^{\tau-1}\rho^t\frac{\Psi_R^L(s_t)}{\sum_{s=0}^{\tau-1}\rho^s})] \text{ since } \varphi \text{ is concave} \\
&\leq \varphi(\widetilde{\mathbb{E}}[\sum_{t=0}^{\tau-1}\rho^t\frac{\Psi_R^L(s_t)}{\sum_{s=0}^{\tau-1}\rho^s}]) \\
&\quad \text{by Jensen inequality, since } \varphi \text{ is concave} \\
&= \varphi(\frac{\mathbb{E}[\sum_{t=0}^{\tau-1}\rho^t\Psi_R^L(s_t) \mid s_0=s]}{\mathbb{E}[\sum_{t=0}^{\tau-1}\rho^t \mid s_0=s]}) \text{ by definition of } \widetilde{\mathbb{E}}.
\end{aligned}$$

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

$$\begin{aligned}
\mu_R^M(s) &= \sup_{\tau>0}\left(\frac{\mathbb{E}[\sum_{t=0}^{\tau-1}\rho^t\Psi_R^M(s_t) \mid s_0=s]}{\mathbb{E}[\sum_{t=0}^{\tau-1}\rho^t \mid s_0=s]}\right) \\
&\leq \sup_{\tau>0}\varphi\left(\frac{\mathbb{E}[\sum_{t=0}^{\tau-1}\rho^t\Psi_R^L(s_t) \mid s_0=s]}{\mathbb{E}[\sum_{t=0}^{\tau-1}\rho^t \mid s_0=s]}\right) \\
&\leq \varphi\left(\sup_{\tau>0}\frac{\mathbb{E}[\sum_{t=0}^{\tau-1}\rho^t\Psi_R^L(s_t) \mid s_0=s]}{\mathbb{E}[\sum_{t=0}^{\tau-1}\rho^t \mid s_0=s]}\right) \text{ since } \varphi \text{ is increasing} \\
&= \varphi(\mu_R^L(s))\,.
\end{aligned}$$

Now, we have by assumption $\mu_C^M(s) = \Psi_C^M \geq \varphi(\Psi_C^L) = \varphi(\mu_C^L(s)$, so that

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

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

3 Risk aversion and optimal strategies

We wish to examine how individual risk aversion modifies dynamics of optimal decisions. Following the Arrow-Pratt definition of absolute risk aversion [9, 7, 4], we say that 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 increasing because U^M and U^L are increasing.

Proposition 2 *Consider two decision-makers, one more risk averse than the other. Assume that, at the beginning, the more risk averse decision-maker selects the random arm. Then, so does the less risk averse decision-maker and, as long as the more risk averse decision-maker selects the random arm, 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 increasing function φ such that $\varphi \circ U^L = U^M$. The state space is here $\mathbb{S} = \mathcal{P}(\mathcal{P}(\mathbb{R}))$, the space of probabilities on the space $\mathcal{P}(\mathbb{R})$ of probabilities on $\mathbb{R}$, and the rewards are given by

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

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

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

The end of the proof follows with Theorem 1 above. □

This Proposition implies that the decision-makers can be ranked by their degree of risk aversion. More risk averse individuals are less likely to select the certain arm in the firts

period (and stick to it). If an invividual is more risk averse than another, he will select for a smaller period of time the random arm. A direct consequence of the above Proposition 2 is that the mean time spent selecting the random arm decreases with the degree of absolute risk aversion.

For risk lovers, concavity is replaced by convexity. Therefore, an invividual more risk lover than another selects for a longer period of time the random arm.

We illustrate graphically in Figure 1 the Proposition 2. In the numerical example, the certain arm has return $w_0 = 40/60$ and the random arm has two returns: $w_- = 38/60$ and $w_+ = 50/60$ ($w_- < w_0 < w_+$). We have used $\rho = 1/1.08$ and the following CARA utility function $V_\theta(x) = -\frac{e^{-\theta x}}{\theta}$. The parameter θ is the Arrow-Pratt degree of absolute risk aversion. The horizontal axis corresponds to the number of w_- and the vertical axis corresponds to the number of w_+ in Figure 1. It shows that the "certain" region (gray zone) gets larger for increasing values of risk aversion, θ, highlighting numerically Proposition 2.

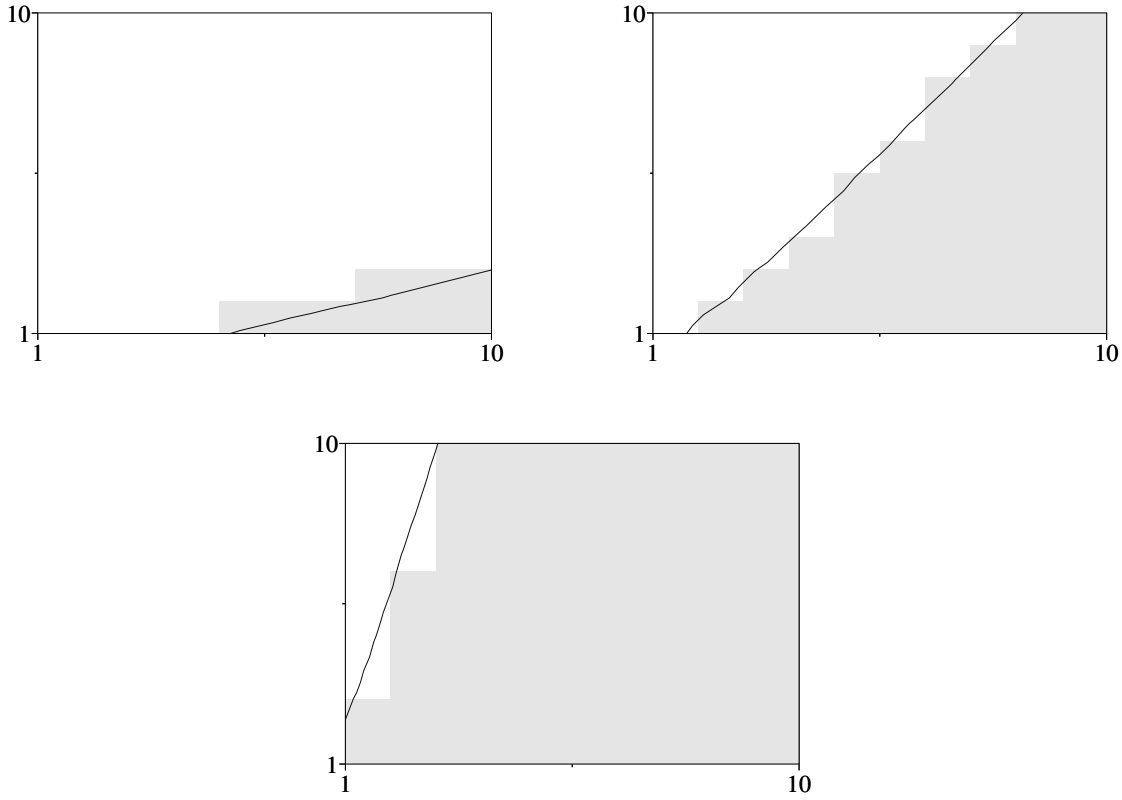

Figure 1: Increasing regions of "certain arm choice" for increasing values of θ (7, 27, 53). The horizontal axis corresponds to the number of low return of the random arm, while the vertical axis corresponds to the high one.

Acknowledgments. We thank the reviewers for pointing to us the literature of operations research on the risk sensitive approach. We also wish to thank the ANR (Agence Nationale de la Recherche) for its support through the network RiskAttitude.

References

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

[2] J.-P. Chancelier, M. De Lara, and A. de Palma. Risk aversion, road choice and the one-armed bandit problem. *Transportation Science*, 41(1):1–14, February 2007.

[3] Eric V. Denardo, Haechurl Park, and Uriel G. Rothblum. Risk-Sensitive and Risk-Neutral Multiarmed Bandits. *Mathematics of Operations Research*, 32(2):374–394, 2007.

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

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

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

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

[8] T. Magnac and J.-M. Robin. Dynamic stochastic dominance in bandit decision problems. *Theory and Decision*, 47:267–295, 1999.

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

[10] M. Rothschild. Searching for the lowest price when the distribution of prices is unknown. *Journal of Political Economy*, 82(4):689–711, 1974.

[11] P. Whittle. *Optimization over Time: Dynamic Programming and Stochastic Control*, volume 1. John Wiley & Sons, New York, 1982.

[12] K. Wolpin. An estimable stochastic model of fertility and child mortality. *Journal of Political Economy*, 92(5):852–874, 1984.