**MACS 30250 (Spring 2019)**

*Perspectives on Computational Research in Economics*

# Literature Review

Submitted by- Nipun Thakurele

**Risk Aversion in Budget Constrained Multi-Armed Bandits**

This literature review is centred around risk aversion in multi-armed bandits. My research question: "Would a more risk-averse decision maker be more willing to take risky actions in case of Budget Constrained Multi-Armed Bandits?" develops on the recent work of Keller et al. (2019) wherein they study risk aversion in a standard two-armed bandit problem. The decision maker has to choose between a safe arm whose payoff is known, and a risky arm that yields unknown payoff. Earlier, Chancelier et al. (2009) have shown the intuitive result: a more risk averse decision maker will be less willing to pull the risky arm. However, Keller et al. (2019) have shown that for a part of the parameter space this intuitive result is overturned. That is, in case of risky arm yielding good payoffs with high frequency, a more risk averse decision maker might be more willing to pull the risky arm.

One of the important contributions of Keller et al. (2019) is the implication of their findings on inferences about risk preferences. Their study asks researchers to be careful while inferring risk preferences from the actions of an agent. In a bandit setup, a more risk averse decision maker might be more willing to take risky actions in comparison to the less risk-averse decision maker. This counter-intuitive result stems from the fact that experimenting with the risky arms in bandit models generates information on the environment. Essentially, the more risk averse decision maker is reducing her/his future risk by showing a greater appetite for risky actions. Their results have applications in the domain of machine learning and reinforcement learning algorithms which process information at high frequency in a multi-armed bandit setup.

I will be using the exponential bandit model with a continuous-time framework following the works of Keller et al. (2019), Keller et. al (2005), Roberts and Weitzman (1981) and Bolton and Harris (1999). My model will be closest to the works of Keller et al. (2019) and Chancelier et al. (2009) who have extended the standard multi-armed bandit model by incorporating risk averse decision maker. As we are considering a continuous-time framework, time $t \in [0, \infty)$ is continuous and we employ an exponential discount rate, $r > 0$. Apart from considering budget constrained problem, the other interesting aspect could have been the use of hyperbolic discounting instead of the exponential discounting used in the current model. Laibson (1997) built upon the previous studies which showed that the individual's discount function are non-exponential (Strotz, 1956). He used the hyperbolic discount functions to analyze the decision-making and preferences of the consumers.

In the standard two-armed bandit problem, the decision maker has to choose between the risky arm R and the safe arm S. The decision maker is provided a unit of 'informational resource' in each time period which can be allocated to either the risky arm or the safe arm. It is known to the decision maker that the safe arm provides lump-sum payoffs of $s > 0$, according to a Poisson process with parameter 1, and in a 'good' state, the risky arm yields the lump-sum payoffs of $h > 0$, according to a Poisson process with parameter $\lambda > 0$. However, there's risk modelled in through type θ of the risky arm which is unknown to the decision maker at $t = 0$. If the arm is 'good' ($\theta = 1$) the pay-off is $h > 0$ according to a Poisson process with parameter $\lambda > 0$ and if the arm is 'bad' ($\theta = 0$), the risky arm pays nothing. Essentially, the decision maker's expected increase in utility is $[(1 - k_t) * u(s) + k_t * p_t * \lambda * u(h)]dt$ where $(1 - k_t) * u(s)$ is her expected pay-off from pulling the safe arm over an interval $[t, t + dt]$. And, she receives $k_t * p_t * \lambda * u(h)$ by using the risky arm over an interval $[t, t + dt]$. Note: $k_t \in \{0, 1\}$ is the 'informational resource' she allocates to the risky arm and $p_t$ captures her belief that the risky arm is of 'good' nature.

The framework presented until now involves the similar methods and solution concepts as in Keller et al. (2005). They have also used the two-armed bandit setup to analyse a game of strategic experimentation. In their study, the players can learn about the state of the risky arm from the action of other players and the "breakthroughs" (exploring that the risky arm is good) occur independently for each player. Unlike our model, the decision maker can allocate fraction of the unit 'information resource' to each arm instead of $\{0, 1\}$. Their model required the decision-maker to be risk-neutral which was relaxed in the studies by Keller et al. (2019) and Chancelier et al. (2009) which aimed at studying how the decision making is influenced by the risk-averseness of a decision maker.

Economists have been studying problems centred around decision making, uncertainty and risk-aversion for a long period of time. There have been developments in the literature in forms of considering expected utilities instead of expected pay-offs, introduction of hyperbolic discounting instead of exponential discounting, etc. One such study which considers the decision making under uncertainty, and with learning is by Chancelier et al. (2009). They measure risk aversion "by the degree of concavity of the utility function V (related to Arrow-

Pratt absolute risk aversion)"[1] and study how risk aversion affects the selection of arms in a multi-armed bandit problem. They found the intuitive result that the more is the decision maker risk-averse (characterized by a more concave utility function), the more is the pulling of the safe arm.

The study by Chancelier et al. (2009) was one of the first research studies to incorporate risk in bandit problems in this particular way. They have also studied decision making by a single individual in a route choice problem (Chancelier et al., 2007). The individual makes a choice between a safe route (involving constant travel time) and a risky route (involving certain travel time). The literature of operations research and computer science mainly studies risk using mean-variance model in which risk is attached with each arm based on the variance in its return, the expected reward. Sani et al. (2012) used the mean-variance model to study the problem of "competing against the arm which had the best risk-return trade-off"[2]. Ding et al. (2013) also uses the variance as a measure of risk but they consider the multi-armed bandit problems which are not only budget-constrained but also have variable costs attached with each arm, that is, the cost of pulling each arm is randomized. These problems have many Internet applications like ad exchange and sponsored search.

The idea of studying budget-limited multi-armed bandit (MAB) problems is not new. Tran-Thanh et al. (2010) introduced the budget-limited multi-armed bandit wherein the decision maker still learns by pulling arms as in standard MAB problem, but now her actions are costly. There is a cost attached with pulling any arm and she is constrained by a fixed budget which makes it a finite horizon problem. Improving upon the first algorithm in their previous work, Tran-Thanh et al. (2012) developed two pulling policies (algorithms) which maximises the agent's total reward within the given budget. It is important to note that the optimal exploitation policy may not be exploring and exploiting the optimal arm repeatedly as a decision maker might only get to pull arms N times, depending upon the cost of pulling each arm and the given budget. However, the study of decision making in budgeted multi-armed bandit problem with risk modelled using expected utility theory is new and will be explored in my paper.

[1] Chancelier, Jean-Philippe, Michel De Lara, and André de Palma. "Risk aversion in expected intertemporal discounted utilities bandit problems." *Theory and decision* 67, no. 4 (2009): 433-440.

[2] Sani, Amir, Alessandro Lazaric, and Rémi Munos. "Risk-aversion in multi-armed bandits." In *Advances in Neural Information Processing Systems*, pp. 3275-3283. 2012.

**References**


Bolton, Patrick, and Christopher Harris. "Strategic experimentation." *Econometrica* 67, no. 2 (1999): 349-374.

Chancelier, Jean-Philippe, Michel De Lara, and André de Palma. "Risk aversion in expected intertemporal discounted utilities bandit problems." *Theory and decision* 67, no. 4 (2009): 433-440.

Chancelier, Jean-Philippe, Michel De Lara, and Andre De Palma. "Risk aversion, road choice, and the one-armed bandit problem." *Transportation Science* 41, no. 1 (2007): 1-14.

Ding, Wenkui, Tao Qin, Xu-Dong Zhang, and Tie-Yan Liu. "Multi-armed bandit with budget constraint and variable costs." In *Twenty-Seventh AAAI Conference on Artificial Intelligence*. 2013.

Keller, Godfrey, Sven Rady, and Martin Cripps. "Strategic experimentation with exponential bandits." *Econometrica* 73, no. 1 (2005): 39-68.

Keller, Godfrey, Vladimír Novák, and Tim Willems. "A note on optimal experimentation under risk aversion." *Journal of Economic Theory* 179 (2019): 476-487.

Laibson, David. "Golden eggs and hyperbolic discounting." *The Quarterly Journal of Economics* 112, no. 2 (1997): 443-478.

Roberts, Kevin, and Martin L. Weitzman. "Funding criteria for research, development, and exploration projects." *Econometrica: Journal of the Econometric Society* (1981): 1261-1288.

Sani, Amir, Alessandro Lazaric, and Rémi Munos. "Risk-aversion in multi-armed bandits." In *Advances in Neural Information Processing Systems*, pp. 3275-3283. 2012.

Strotz, Robert Henry. "Myopia and inconsistency in dynamic utility maximization." *The Review of Economic Studies* 23, no. 3 (1955): 165-180.

Tran-Thanh, Long, Archie Chapman, Alex Rogers, and Nicholas R. Jennings. "Knapsack based optimal policies for budget–limited multi–armed bandits." In *Twenty-Sixth AAAI Conference on Artificial Intelligence*. 2012.

Tran-Thanh, Long, Archie Chapman, Enrique Munoz de Cote, Alex Rogers, and Nicholas R. Jennings. "Epsilon–first policies for budget–limited multi-armed bandits." In *Twenty-Fourth AAAI Conference on Artificial Intelligence*. 2010.