Risk Aversion in Budget Constrained Multi-Armed Bandits

Research Question:
Would a more risk-averse decision maker be more willing to take risky actions in case of Budget Constrained Multi-Armed Bandits?

Submitted by-
Nipun Thakurele

What is Multi-Armed Bandit

It is a sequential experiment with the goal of achieving the largest possible reward from a payoff distribution with unknown parameters.

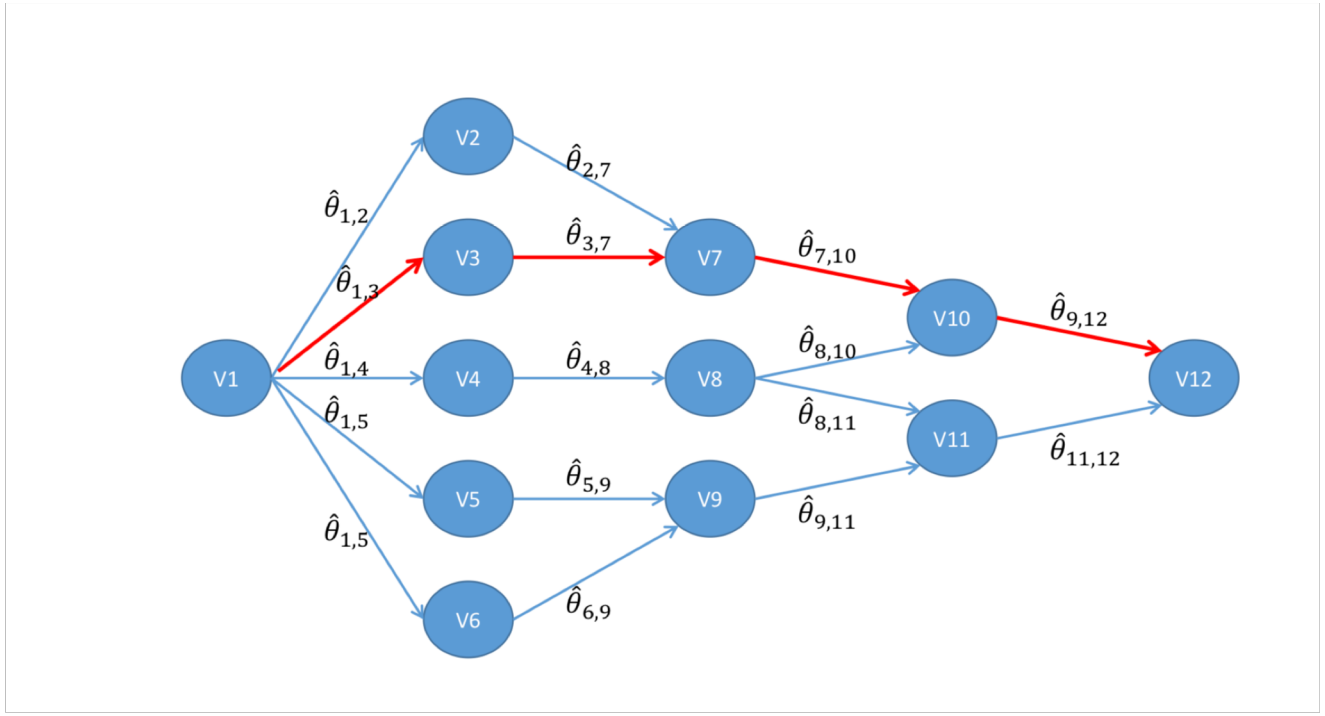
At each stage, the experimenter must decide which arm of the experiment to observe next.

The choice involves a fundamental trade-off between the utility gain from *exploiting* arms that appear to be doing well (based on limited sample information) vs *exploring* arms that might potentially be optimal, but which appear to be inferior because of sampling variability.

The trade-off has also been referred to as ‘earn vs learn.’

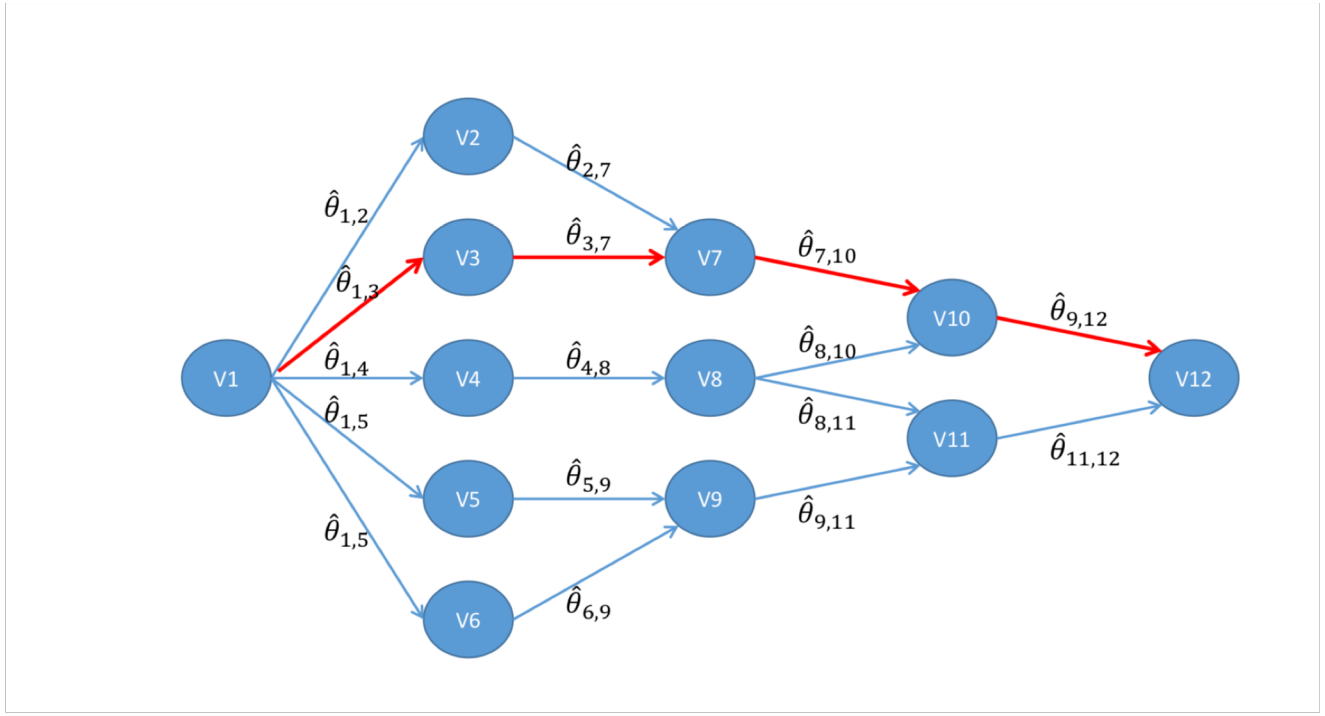
- Example:Online Shortest Path

- Alice needs to commute from home to college.
- She would like to use the path that takes the least average travel time, but there's uncertainty involved with the
- Travel time along different paths. How can she learn and minimize her total travel time?
- Let's formulate this as a shortest path problem on a graph G = (V,E)



- Example:Online Shortest Path

- Alice needs to commute from home to college.
- She would like to use the path that takes the least average travel time, but there's uncertainty involved with the
- Travel time along different paths. How can she learn and minimize her total travel time?
- Let's formulate this as a shortest path problem on a graph G = (V,E)



Importance in Economics

Experimentation and Matching:
Application of the bandit framework to learning in matching markets
Example: labor and consumer good markets.
Consider a competitive labor market wherein a worker chooses employment in one of the K firms
Her (random) productivity in firm k is parametrized by a real variable θ^k

The bandit problem provides a framework to the study of learning about the match specific productivities.
Over time, a worker's productivity in a specific job becomes know more precisely.
In the event of a poor match, separation occurs in equilibrium and job turnover arises as natural by-product of the learning process.
On the other hand, over time the likelihood of separation eventually decreases,
as conditional on being still on the job, the likelihood of a good match increases.

Importance in Economics

Experimentation and Matching:
Application of the bandit framework to learning in matching markets
Example: labor and consumer good markets.
Consider a competitive labor market wherein a worker chooses employment in one of the K firms
Her (random) productivity in firm k is parametrized by a real variable θ^k

The bandit problem provides a framework to the study of learning about the match specific productivities.
Over time, a worker's productivity in a specific job becomes know more precisely.
In the event of a poor match, separation occurs in equilibrium and job turnover arises as natural by-product of the learning process.
On the other hand, over time the likelihood of separation eventually decreases,
as conditional on being still on the job, the likelihood of a good match increases.

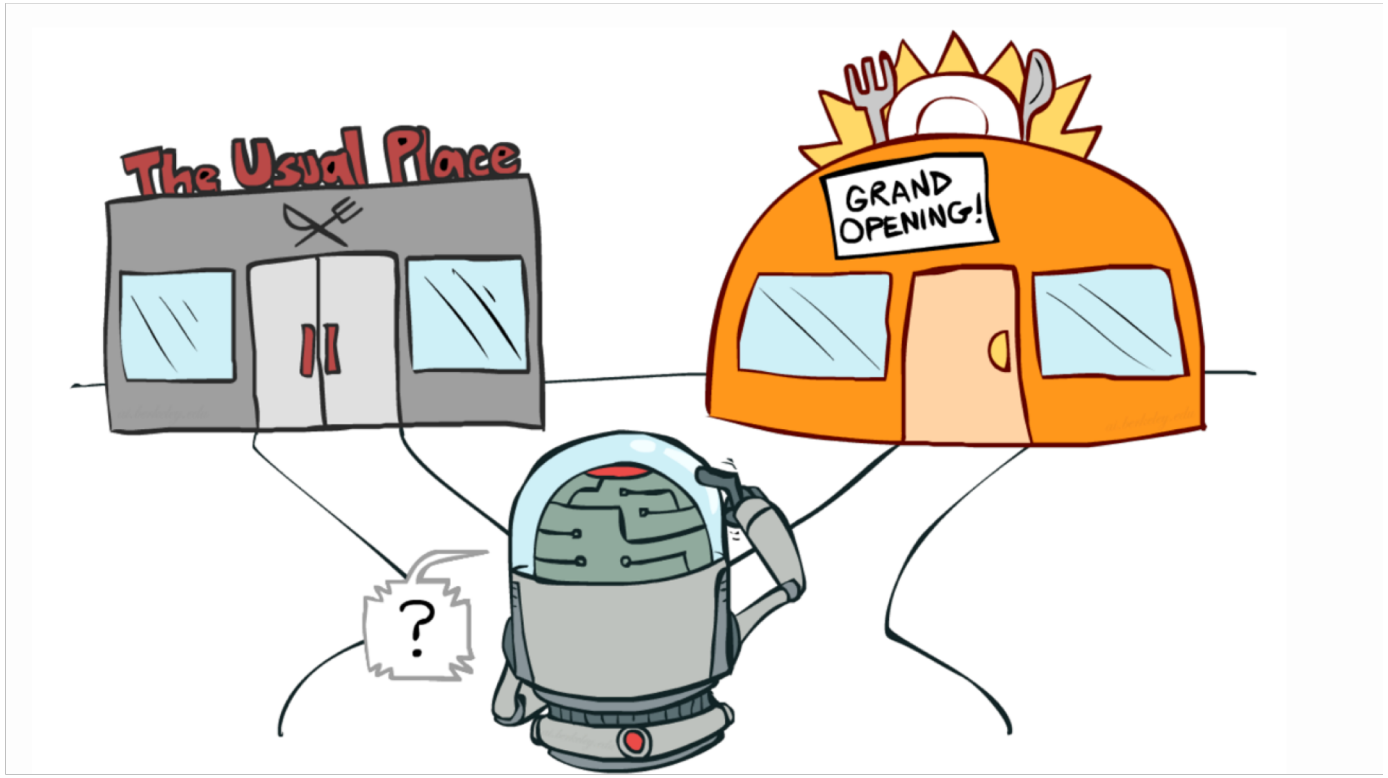
Model

- Time $t \in [0, \infty)$ is continuous and the discount rate is $r > 0$.
- The player is facing a two-armed bandit problem, and at time t can select between the risk arm, R or safe arm, S.
- S provides lump-sum payoffs of $s > 0$ (with the value of s fixed and known to the player)
- The payoffs are generated according to a Poisson process with parameter 1 (which is also known).
- Risky arm's type θ, and hence the size of its payoff, is unknown to the agent at $t = 0$.
- She knows that the arm is either 'good' ($\theta = 1$) or 'bad' ($\theta = 0$).
- At time t, the player holds a belief p_t that the risky arm is good.
- A good arm yields lump-sum payoffs of h according to a Poisson process with parameter $\lambda > 0$
- A bad arm pays zero
- The expected increase in her utility is $[(1 - k_t)u(s) + k_t p_t \lambda u(h)]dt$

Model

- Time $t \in [0, \infty)$ is continuous and the discount rate is $r > 0$.
- The player is facing a two-armed bandit problem, and at time t can select between the risk arm, R or safe arm, S.
- S provides lump-sum payoffs of $s > 0$ (with the value of s fixed and known to the player)
- The payoffs are generated according to a Poisson process with parameter 1 (which is also known).
- Risky arm's type θ, and hence the size of its payoff, is unknown to the agent at $t = 0$.
- She knows that the arm is either 'good' ($\theta = 1$) or 'bad' ($\theta = 0$).
- At time t, the player holds a belief p_t that the risky arm is good.
- A good arm yields lump-sum payoffs of h according to a Poisson process with parameter $\lambda > 0$
- A bad arm pays zero
- The expected increase in her utility is $[(1 - k_t)u(s) + k_t p_t \lambda u(h)]dt$

Contribution



Contribution

- Specifically, given a bandit with N distinct arms, each arm indexed by i ∈ [N] is associated with an unknown reward and cost distribution with unknown means.
- Realizations of costs and rewards are independently and identically distributed.
- At each round t, the decision maker plays exactly 1 arm and subsequently observes the individual costs and rewards only for the played arm
- Before the game starts, the player is given a budget $0 < B \in R+$ to pay for the materialized costs based on the arm played.
- The game stops when the budget is exhausted