Bandit Problems with Lévy Payoff Processes

Eilon Solan

Tel Aviv University

Joint with Asaf Cohen

Multi-Arm Bandits

A single player sequential decision making problem.
Time is continuous or discrete.

The decision maker has finitely many actions (called “arms”).
Each arm **i** generates a payoff which is a stochastic process $(x^i(s))$.

The distribution of $(x^i(s))$ is not known. It is drawn at the outset from a known family of distributions according to a known probability distribution p^i.

At every time **t**, the decision maker has to choose one arm **i**, and he observes the next realization of the process $(x^i(s))$ (at the s^{th} time in which he chooses arm **i** he observes $x^i(s)$).

The goal: to maximize the discounted payoff.

There is a tradeoff between exploration and exploitation.

Multi-Arm Bandits: cont.

A strategy: an indication of which arm to choose after any history (can use randomization).

Question: What is the structure of the optimal strategy?

Answer: The optimal strategy is an index strategy.
For every arm **i**, calculate an **index**, which depends only on past observations of that arm. Choose the arm with maximal index.

The index is the unique real number **s**, such that in the problem that contains two arms, a safe arm that always generates **s**, and arm **i** (given past observations on the arm), the decision maker is indifferent between the arms at time 0.

Gittins and Jones (1979) for discrete time problems.
Karatzas (1985) and Kaspi and Mandelbaum (1995) for special continuous time problems, when the distribution of each arm is **known**.

One-Arm Bandits with Two Types

One safe arm that always generates **s**.
One risky arm that generates a stochastic process **(X(t))**.

The risky arm can have two types: **High** or **Low**.
The **High** type's expected "payoff per time" is higher than **s**,
The **Low** type's expected "payoff per time" is lower than **s**.

p(t) = the belief at time **t** that the type of the risky arm is High.

When time is discrete: optimal strategy is a cut-off strategy:
Choose the risky arm as long as $p(t) > p^*$.
Switch to the safe arm once $p(t) \leq p^*$.
The same if time is continuous and **(X(t))** is a Brownian motion with drift.

Question: Can we provide an explicit expression to p^* ?

Lévy Payoff Processes

The risky arm generates a stochastic payoff $(X_h(t))$ or $(X_l(t))$.

The Decision Maker has to decide, for every time interval **[t,t+dt)**, the proportion **k** of the time interval to invest in the risky arm (the rest will be invested in the safe arm).

The payoff at the time interval **[t,t+dt)**, is

$$\underbrace{k\,dt \times s}_{\text{safe arm}} + \underbrace{(1-k) \times dX(t)}_{\text{risky arm}}$$

Payoff is discounted at a discount rate **r**.

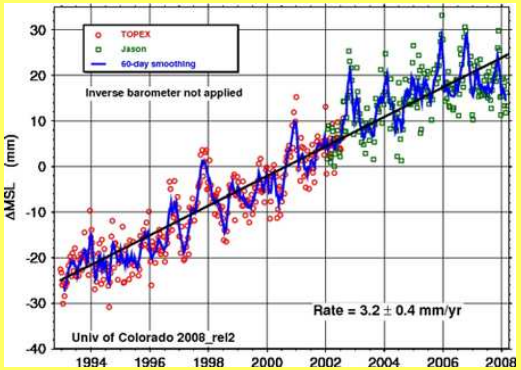
Both $(X_h(t))$ and $(X_l(t))$ are Lévy processes, with different parameters.

A Lévy Process

The continuous time analog of random walk.

A Lévy process is a continuous time process **(X(t))** that satisfies:

1) It starts at the origin: **X(0) = 0**.
2) It has stationary independent increments.
3) It has Càdlàg modification (RCLL): continuous from the right, has limits from the left.



Special cases:

1) Brownian motion.
2) Linear drift: **X(t) = μt**.
3) Poisson process: a lump sum **c** arrives at a time that has a Poisson distribution.
4) Compound Poisson process: the sum of Poisson processes (there are many possible jumps. The expected number of jumps with size in a Borel subset **A** of **R** up to time 1 is **ν(A)**).
5) The sum of independent copies of the above.

A Lévy Process, cont.

The Lévy-Ito decomposition theorem: A Lévy process with finite Lévy measure is the sum of a linear drift, a Brownian motion, and a compound Poisson process (independent of the Brownian motion).

The Lévy measure is finite if $\mathbf{E[\nu] := \int x\, \nu(dx)}$ is finite.
This is the expected jump size per time in [t,t+dt).

I will present results for Lévy processes with finite Lévy measure, but they were extended to Lévy processes with infinite Lévy measure.

Back to the Bandit Problem

	High type	Low type
Linear drift	μ_h	μ_l
Standard deviation of Brownian motion	σ_h	σ_l
Intensity of jump's size	ν_h	ν_l

Back to the Bandit Problem

	High type	Low type
Linear drift	μ_h	μ_l
Standard deviation of Brownian motion	$\boldsymbol{\sigma_h}$	$\boldsymbol{\sigma_l}$
Intensity of jump's size	ν_h	ν_l

If the standard deviations of the Brownian motions differ, the DM can identify the type of the arm at an infinitesimal time interval.

Back to the Bandit Problem

	High type	Low type
Linear drift	μ_h	μ_l
Standard deviation of Brownian motion	σ	σ
Intensity of jump's size	ν_h	ν_l

Back to the Bandit Problem

	High type	Low type
Linear drift	μ_h	μ_l
Standard deviation of Brownian motion	σ	σ
Intensity of jump's size	ν_h	ν_l

If, e.g., a jump that has 0 probability under ν_l and positive probability under ν_h arrives, then the DM deduces that the arm is High.

Assumptions

	High type	Low type
Linear drift	μ_h	μ_l
Standard deviation of Brownian motion	σ	σ
Intensity of jump's size	ν_h	ν_l

High type is better than **Low** type in a strong sense:

1) **High** type is better than safe arm is better than **Low** type:

$$\mu_h + E[\nu_h] > s > \mu_l + E[\nu_l].$$

2) Jumps of **High** type dominate jumps of **Low** type: $\nu_h(A) \geq \nu_l(A)$ for every Borel set **A**.

Implications of Assumptions

High type is better than **Low** type in a strong sense:

1) **High** type is better than safe arm is better than **Low** type:

$$\mu_h + E[\nu_h] > s > \mu_l + E[\nu_l].$$

2) Jumps of **High** type dominate jumps of **Low** type: $\nu_h(A) \geq \nu_l(A)$ for every Borel set **A**.

Suppose that a jump of size **x** occurs:

If $\nu_l(x) = 0$, then the DM known that the type is **High**.

If $\nu_l(x) > 0$, then the probability of the **High** type does not decrease.

If there is a jump, the probability of the **High** type does not decrease.

The Dynamic Programming Principle

Write the dynamic programming equation of the optimal payoff (Bolton and Harris (1999), Keller, Rady and Cripps (2005):

$$U(p) = \max\{\, s, \underbrace{(p(E[\nu_h]+\mu_h) + (1-p)(E[\nu_l]+\mu_l))rdt}_{\text{instanteous payoff}} + \underbrace{\exp(-rdt)\, E[U(p+dp)]}_{\text{continuation payoff}} \}.$$

U is continuous, non-decreasing and convex.
Use Taylor expansion (second order), and obtain a differential equation (with U' and U'').
Find its solution.

The Solution

Define:

$$f(\eta) = \int \nu_l(dx)\left(\frac{\nu_l(dx)}{\nu_h(dx)}\right)^{\eta} + \eta(E[\nu_h]-E[\nu_l]) - E[\nu_l] + \tfrac{1}{2}(\eta+1)\eta\left(\frac{\mu_h-\mu_l}{\sigma}\right)^2 - r.$$

The equation $f(\eta)=0$ has a unique solution α in the interval $(0,\infty)$.

The optimal cut-off is: $p^* := \dfrac{\alpha(s-g_l)}{\alpha(g_h-g_l) + (g_h-s)}$

Here:
$g_h = \mu_h + E[\nu_h]$,
$g_l = \mu_l + E[\nu_l]$.

The optimal payoff is:

$$U(p) = \begin{cases} s, & \text{if } p \le p^* \\ g_l + (g_h-g_l)p + C(1-p)\left(\dfrac{1-p}{p}\right)^{\alpha} \end{cases}$$

$$C := \frac{s-g_l - p^*(g_h-g_l)}{(1-p^*)\left(\frac{1-p^*}{p^*}\right)^{\alpha}}$$

Application 1: Pricing Information

Suppose that $\mathbf{X_h(t) = Y_h(t) + Z_h(t)}$, and $\mathbf{X_l(t) = Y_l(t) + Z_l(t)}$.

The DM observes only $\mathbf{Y_h}$ (or $\mathbf{Y_l}$), but his payoff also depends on $\mathbf{Z_h}$ (or $\mathbf{Z_l}$).

The DM can observe $\mathbf{Z_h}$ (or $\mathbf{Z_l}$) at a cost $\mathbf{c}$.

Question: What is the optimal strategy in each case?
Question: What is the fair price $\mathbf{c}$ of the additional information?

Application 2: Optimism versus Pessimism

Suppose that the DM believes that the initial prior is q_0, even though the true probability of the High type is p_0.

The DM uses the optimal strategy given q_0.
Question: What is the expected payoff?

Question: Who will fare better, an optimist with a prior belief $p_0 + \varepsilon$, or a pessimist with a prior belief $p_0 - \varepsilon$?