Multi-armed bandits under uncertainty aversion

Tanut Treetatanthiploet

Mathematical Institute
University of Oxford

Based on joint work with Prof. Samuel N. Cohen

12th European Summer School in Financial Mathematics
Padova, September 2019

UNIVERSITY OF OXFORD
Mathematical Institute

Oxford Mathematics

Toy Example

Gambling is a way of buying hope on credit. – Alan Wykes

Suppose there are two biased coins.
You will gain £1 for a head

- 1st coin: 3 tosses with 2 heads.
- 2nd coin: 3000 tosses with 2000 heads.

Toy Example

Gambling is a way of buying hope on credit. – Alan Wykes

Suppose there are two biased coins.
You will gain £1 for a head

- 1st coin: 3 tosses with 2 heads.
- 2nd coin: 3000 tosses with 2000 heads.

- We are biased toward a less uncertain choice.

Toy Example

Gambling is a way of buying hope on credit. – Alan Wykes

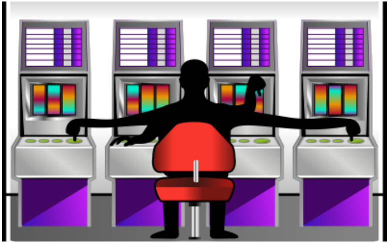
Suppose there are two biased coins.
You will gain £1 for a head

- 1st coin: 3 tosses with 2 heads.
- 2nd coin: 3000 tosses with 2000 heads.

- We are biased toward a less uncertain choice.
- What shall we do if we need to repeat for a million tosses?

Multi-armed bandits problem

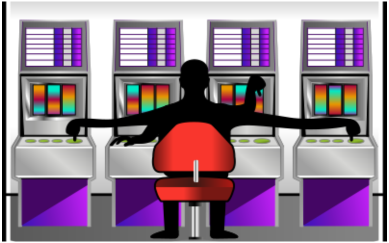
Gambling is a way of buying hope on credit. – Alan Wykes



- Suppose there are M slot machines.
- One machine can be played at a time.
- Each machine may have its own state.
- The machines are independent.
- Playing a machine generates a cost and its state may evolve.

Multi-armed bandits problem

Gambling is a way of buying hope on credit. – Alan Wykes



- Suppose there are M slot machines.
- One machine can be played at a time.
- Each machine may have its own state.
- The machines are independent.
- Playing a machine generates a cost and its state may evolve.

	Distribution	Evolving state
Risky bandit	Known	Yes
Stationary bandit	Unknown	No
Non-stationary bandit	Unknown	Yes

Gittins' index theorem as a risk model

An optimist is a guy that has never had much experience. –Don Marquis

University of Oxford
Mathematical Institute

Consider a risky bandit problem with known distribution.

- Costs $(h^{(m)}(t))$ are not IID.
- Objective: Minimise $\mathbb{E}\Big(\sum_{n=0}^{\infty} \beta^n h^{(\rho_n)}(t_n^{\rho}) \Big)$.

There exist indices associated to each machine which can be evaluated independently such that the optimal policy is to play at each epoch a machine of the lowest index $\gamma^{(m)}$. (Gittins, 1979)

Gittins' index theorem as a risk model

An optimist is a guy that has never had much experience. –Don Marquis

UNIVERSITY OF OXFORD
Mathematical Institute

Consider a risky bandit problem with known distribution.

- Costs $(h^{(m)}(t))$ are not IID.
- Objective: Minimise $\mathbb{E}\Big(\sum_{n=0}^{\infty} \beta^n h^{(\rho_n)}(t_n^{\rho}) \Big)$.

There exist indices associated to each machine which can be evaluated independently such that the optimal policy is to play at each epoch a machine of the lowest index $\gamma^{(m)}$. (Gittins, 1979)

By modeling costs $h^{(m)}(t)$ under a Bayesian perspective, we can show that

$$\gamma^{(m)} \approx \bar{x}^{(m)} - \sigma^{(m)} \psi\left(\frac{1}{(n^{(m)}+1)(1-\beta)} \right) \qquad \text{(Brezzi and Lai, 2002)}$$

where $\bar{x}^{(m)}$ and $\sigma^{(m)}$ are posterior mean and s.d. and ψ is positive and nondecreasing.

Optimistic Analogy for Reinforcement learning

An optimist is a guy that has never had much experience. –Don Marquis

Mathematical Institute

For stationary bandit problem,

- The m^{th} machine generates IID costs $\left(h^{(m)}(t)\right)_{t\in\mathbb{N}}$ with mean $\mu^{(m)}$.

$$\rho^* = \arg\min_m \left(\text{Est. of } \mu^{(m)} - \underbrace{\text{Learning Premium}}_{\text{decreases in } n^{(m)}}\right).$$

i.e. We have less learning reward when we are more certain about our estimator.

Uncertainty Aversion

If you expect the worst, you'll never be disappointed – Sarah Dessen

Suppose we will only play once.

- This is equivalent to setting $\beta = 0$.
- Gittins' objective: Minimise $\mathbb{E}\Big(h^{(\rho_0)}(t_0^\rho)\Big) = \bar{x}^{(\rho_0)}$.

Uncertainty Aversion

If you expect the worst, you'll never be disappointed – Sarah Dessen

Suppose we will only play once.

- This is equivalent to setting $\beta = 0$.
- Gittins' objective: Minimise $\mathbb{E}\Big(h^{(\rho_0)}(t_0^{\rho})\Big) = \bar{x}^{(\rho_0)}$.

No accounting for uncertainty!

Uncertainty Aversion

If you expect the worst, you'll never be disappointed – Sarah Dessen

Suppose we will only play once.

- This is equivalent to setting $\beta = 0$.
- Gittins' objective: Minimise $\mathbb{E}\Big(h^{(\rho_0)}(t_0^{\rho})\Big) = \bar{x}^{(\rho_0)}$.

No accounting for uncertainty!

Including uncertainty aversion, we expect to have

$$\gamma^{(m)} = \text{Est. of } \mu^{(m)} + \Big(\underbrace{-\text{Learning premium} + \text{Uncertainty aversion}}_{\text{Uncertainty valuation}}\Big).$$

Classical control problem under uncertainty aversion:

$$\inf_{\rho} \sup_{\mathbb{Q}} \mathbb{E}^{\mathbb{Q}}\Big(\sum_{n=1}^{L} \beta^n h^{(\rho_n)}(t_n^{\rho}) \Big) =: \inf_{\rho} \mathcal{E}\Big(\sum_{n=1}^{L} \beta^n h^{(\rho_n)}(t_n^{\rho}) \Big).$$

Time-consistent Nonlinear Expectation

If you expect the worst, you'll never be disappointed. – Sarah Dessen

Classical control problem under uncertainty aversion:

$$\inf_{\rho}\sup_{\mathbb{Q}}\mathbb{E}^{\mathbb{Q}}\Big(\sum_{n=1}^{L}\beta^n h^{(\rho_n)}(t_n^{\rho})\Big) =: \inf_{\rho}\mathcal{E}\Big(\sum_{n=1}^{L}\beta^n h^{(\rho_n)}(t_n^{\rho})\Big).$$

We say a system of operator $\mathcal{E}\Big(\cdot\Big|\mathcal{F}_t\Big): L^\infty(\mathbb{P},\mathcal{F}_T)\to L^\infty(\mathbb{P},\mathcal{F}_t): t=0,1,..,T$ is an $(\mathcal{F}_t)$-consistent coherent nonlinear expectation if it satisfies strict monotonicity, positive homogeneity, subadditivity and Lebesgue property (lower semi-continuity) and

- $(\mathcal{F}_t)$-consistency: for $t\le t'\le T$,

$$\mathcal{E}\Big(\mathcal{E}(X|\mathcal{F}_{t'})\Big|\mathcal{F}_t\Big) = \mathcal{E}(X|\mathcal{F}_t).$$

Classical control problem under uncertainty aversion:

$$\inf_{\rho} \sup_{\mathbb{Q}} \mathbb{E}^{\mathbb{Q}}\Big(\sum_{n=1}^{L} \beta^n h^{(\rho_n)}(t_n^\rho)\Big) =: \inf_{\rho} \mathcal{E}\Big(\sum_{n=1}^{L} \beta^n h^{(\rho_n)}(t_n^\rho)\Big).$$

We say a system of operator $\mathcal{E}\Big(\cdot \Big| \mathcal{F}_t\Big) : L^\infty(\mathbb{P}, \mathcal{F}_T) \to L^\infty(\mathbb{P}, \mathcal{F}_t) : t = 0, 1, .., T$ is an $(\mathcal{F}_t)$-consistent coherent nonlinear expectation if it satisfies strict monotonicity, positive homogeneity, subadditivity and Lebesgue property (lower semi-continuity) and

- $(\mathcal{F}_t)$-consistency: for $t \le t' \le T$,

$$\mathcal{E}\Big(\mathcal{E}(X|\mathcal{F}_{t'})\Big|\mathcal{F}_t\Big) = \mathcal{E}(X|\mathcal{F}_t).$$

The filtration $(\mathcal{F}_t)$ must be identified in advance.

Our information structures

If you expect the worst, you'll never be disappointed. – Sarah Dessen

We have M bandits, each with a filtered space $(\Omega^{(m)}, (\mathcal{F}_t^{(m)}), \mathbb{P}^{(m)})$ and a consistent coherent nonlinear expectation

$$\mathcal{E}^{(m)}\left(X \middle| \mathcal{F}_t^{(m)}\right) = \operatorname*{ess\,sup}_{\mathbb{Q}\in\mathcal{Q}^{(m)}} \mathbb{E}^{\mathbb{Q}}\left(X \middle| \mathcal{F}_t^{(m)}\right)$$

(See Follmer & Schied, 2016).

- Define the *orthant space* by $\bar{\Omega} = \bigotimes_m \Omega^{(m)}$, similarly $\bar{\mathbb{P}}$, and

$$\bar{\mathcal{F}}(\underline{s}) = \bigotimes_m \mathcal{F}^{(m)}(s^{(m)}) \;:\; \underline{s} = (s^{(1)}, s^{(2)}, \dots, s^{(m)}).$$

- Define the orthant nonlinear expectation by

$$\mathfrak{E}_s\left(Y\right) = \operatorname*{ess\,sup}_{\mathbb{Q}\in\bar{\mathcal{Q}}} \mathbb{E}^{\mathbb{Q}}\left(Y \middle| \bar{\mathcal{F}}(\underline{s})\right) \;:\; \bar{\mathcal{Q}} = \{\mathbb{Q} = \bigotimes_m \mathbb{Q}^{(m)} \text{ for } \mathbb{Q}^{(m)} \in \mathcal{Q}^{(m)}\}.$$

Time consistency

When the Facts Change, I Change My Mind. What Do You Do, Sir?–Keynes

One may want to minimise

$$\mathfrak{E}_0\Big(\sum_{n=1}^{L}\beta^n h^{(\rho_n)}(t_n^\rho)\Big) =: \mathfrak{E}_0\big(H^\rho\big)$$

but $\mathfrak{E}_s$ does not satisfy time-consistency.

One may want to minimise

$$\mathfrak{E}_0\Big(\sum_{n=1}^{L}\beta^n h^{(\rho_n)}(t_n^\rho)\Big) =: \mathfrak{E}_0\big(H^\rho\big)$$

but $\mathfrak{E}_s$ does not satisfy time-consistency.

- No DPP $\Rightarrow$ Curse of dimensionality.

One may want to minimise

$$\mathfrak{E}_0\Big(\sum_{n=1}^{L} \beta^n h^{(\rho_n)}(t_n^\rho)\Big) =: \mathfrak{E}_0\big(H^\rho\big)$$

but $\mathfrak{E}_s$ does not satisfy time-consistency.

- No DPP $\Rightarrow$ Curse of dimensionality.
- Inconsistency in decision making. i.e. ‘Optimal’ strategies may not be followed in the future.

Time consistency

When the Facts Change, I Change My Mind. What Do You Do, Sir?–Keynes

One may want to minimise

$$\mathfrak{E}_0\Big(\sum_{n=1}^{L}\beta^n h^{(\rho_n)}(t_n^\rho)\Big) =: \mathfrak{E}_0\big(H^\rho\big)$$

but $\mathfrak{E}_s$ does not satisfy time-consistency.

- No DPP $\Rightarrow$ Curse of dimensionality.
- Inconsistency in decision making. i.e. 'Optimal' strategies may not be followed in the future.
- Using an indifference valuation perspective can be helpful.

‘Gittins’ optimality’

When the Facts Change, I Change My Mind. What Do You Do, Sir?–Keynes

- $\mathfrak{E}_0\big(H^\rho - \mathfrak{E}_0\big(H^\rho\big)\big) = 0.$
- $\min_\rho \mathfrak{E}_0\big(H^\rho\big) := C^{\rho^*} \leq C^\rho$ where $C^\rho \in \mathbb{R}$ and $\mathfrak{E}_0\big(H^\rho - C^\rho\big) = 0.$

Mathematical Institute

- $\mathfrak{E}_0\big(H^\rho - \mathfrak{E}_0\big(H^\rho\big)\big) = 0.$
- $\min_\rho \mathfrak{E}_0\big(H^\rho\big) := C^{\rho^*} \leq C^\rho$ where $C^\rho \in \mathbb{R}$ and $\mathfrak{E}_0\big(H^\rho - C^\rho\big) = 0.$

Definition

We say Y^ρ is a compensator of a cost H^ρ (under strategy ρ) if

$$\mathfrak{E}_0\big(H^\rho - Y^\rho\big) = 0.$$

We say ρ^* is a Gittins' optimum if there exists a compensator family $\{Y^\rho\}_\rho$ such that $Y^{\rho^*} \leq Y^\rho$.

- In a dynamic setting, we require Y^ρ to be predictable and supercompensate at all times.

For a machine m, we have $(\mathcal{F}_t^{(m)})_{t\geq 0}$-adapted random costs $h^{(m)}(t)$.

Theorem

A Gittins' optimal strategy for the cost $H^\rho := \sum_{n=1}^{L} \beta^n h^{(\rho_n)}(t_n^\rho)$ can be given by always playing a machine with the minimal index $\gamma^{(m)}(s)$ where

$$\gamma^{(m)}(s) := \operatorname{ess\,inf}\left\{\gamma : \operatorname*{ess\,inf}_{\tau\in\mathcal{T}^{(m)}(s)} \mathcal{E}^{(m)}\left(\sum_{t=1}^{\tau} \beta^t\big(h^{(m)}(s+t)-\gamma\big)\,\middle|\, \mathcal{F}_s^{(m)}\right)\leq 0\right\}.$$

The main theorem

There is no more miserable human being than one in whom nothing is habitual but indecision. —W. James

For a machine m, we have $(\mathcal{F}_t^{(m)})_{t\geq 0}$-adapted random costs $h^{(m)}(t)$.

Theorem

A Gittins' optimal strategy for the cost $H^\rho := \sum_{n=1}^L \beta^n h^{(\rho_n)}(t_n^\rho)$ can be given by always playing a machine with the minimal index $\gamma^{(m)}(s)$ where

$$\gamma^{(m)}(s) := \text{ess inf}\left\{\gamma : \operatorname*{ess\,inf}_{\tau\in\mathcal{T}^{(m)}(s)} \mathcal{E}^{(m)}\left(\sum_{t=1}^{\tau}\beta^t\big(h^{(m)}(s+t)-\gamma\big)\,\middle|\,\mathcal{F}_s^{(m)}\right)\leq 0\right\}.$$

- $\gamma^{(m)}$ can be found by solving a nonlinear optimal stopping problem independently for each machine.

The main theorem

There is no more miserable human being than one in whom nothing is habitual but indecision. —W. James

Mathematical Institute

For a machine m, we have $(\mathcal{F}_t^{(m)})_{t\geq 0}$-adapted random costs $h^{(m)}(t)$.

Theorem

A Gittins' optimal strategy for the cost $H^\rho := \sum_{n=1}^{L} \beta^n h^{(\rho_n)}(t_n^\rho)$ can be given by always playing a machine with the minimal index $\gamma^{(m)}(s)$ where

$$\gamma^{(m)}(s) := \operatorname{ess\,inf}\left\{\gamma : \operatorname*{ess\,inf}_{\tau\in\mathcal{T}^{(m)}(s)} \mathcal{E}^{(m)}\left(\sum_{t=1}^{\tau}\beta^t\big(h^{(m)}(s+t)-\gamma\big)\,\middle|\,\mathcal{F}_s^{(m)}\right)\leq 0\right\}.$$

- $\gamma^{(m)}$ can be found by solving a nonlinear optimal stopping problem independently for each machine.
- Playing based on indices yields a consistent decision. (i.e. an optimal decision follows through.)

Bernoulli Bandit

But to us, probability is the very guide of life. —Joseph Butler

UNIVERSITY OF OXFORD
Mathematical Institute

- Bernoulli bandit: $h(t) \sim B(1, \theta)$: θ is unknown.
- We model the uncertainty by

$$\mathcal{E}^k_{(t)}(\cdot) := \operatorname*{ess\,sup}_{\theta \in \Theta^k_t} \mathbb{E}^\theta(\cdot)$$

where Θ^k_t is a posterior credible set of size k (under an improper prior).

- We extend it by $\mathcal{E}^k\left(\cdot \middle| \mathcal{F}_t\right) := \mathcal{E}^k_{(t)}\left(\ldots \mathcal{E}^k_{(T-1)}(\cdot) \ldots\right)$.

Bernoulli Bandit

But to us, probability is the very guide of life. —Joseph Butler

- Bernoulli bandit: $h(t) \sim B(1, \theta)$: θ is unknown.
- We model the uncertainty by

$$\mathcal{E}^k_{(t)}(\cdot) := \operatorname*{ess\,sup}_{\theta \in \Theta^k_t} \mathbb{E}^\theta(\cdot)$$

where Θ^k_t is a posterior credible set of size k (under an improper prior).

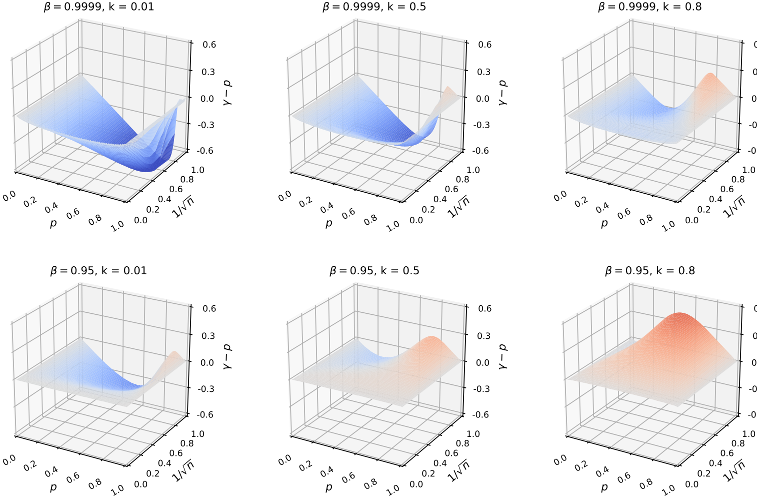
- We extend it by $\mathcal{E}^k\big(\cdot\,\big|\mathcal{F}_t\big) := \mathcal{E}^k_{(t)}\big(\ldots\mathcal{E}^k_{(T-1)}(\cdot)\ldots\big)$.
- It can be shown that

$$\gamma(t) = \gamma_{k,\beta}\left(p_t, \frac{1}{\sqrt{n_t}}\right) =: p_t + \text{Uncertainty valuation.}$$

Difference between γ and p (Uncertainty valuation)

But to us, probability is the very guide of life. —Joseph Butler

UNIVERSITY OF OXFORD
Mathematical Institute



Monte-Carlo Simulation

Facts are stubborn things, but statistics are pliable.—Mark Twain

- Randomly choose a and b independently from $\Gamma(1, 1/100)$.
- Take 50 samples from $\mathrm{Beta}(a, b)$ and use them as true probabilities of the 50 Bernoulli bandits.
- Evaluate algorithms over 10000 trials/simulation.
- Start with initial information of size 10 from each bandit.

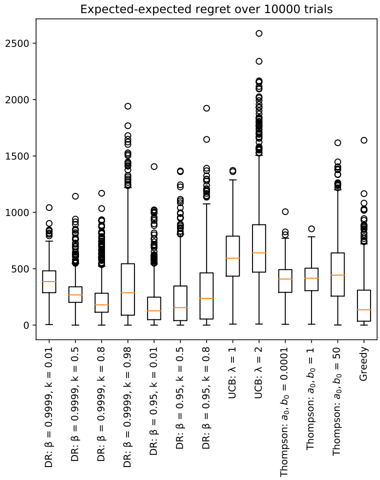
We will consider the performance of each algorithm by considering

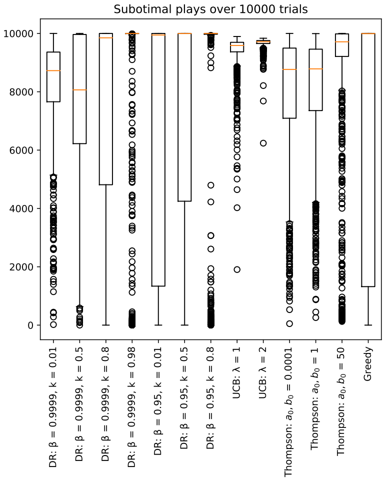
- Expected-expected regret: $R(L) := \sum_{n=1}^{L} \left(\theta^{(\rho_n)} - \theta^*\right)$.
- Suboptimal plays: $N_\vee(L) := \sum_{n=1}^{L} \mathbb{I}\left(\theta^{(\rho_n)} \neq \theta^*\right)$.

Performance

Facts are stubborn things, but statistics are pliable.—Mark Twain

UNIVERSITY OF OXFORD
Mathematical Institute





Conclusion

The finest studies are like the finest of anything else: They cost big bucks. —Charles Wheelan

University of Oxford Mathematical Institute

Contribution

- We propose an alternative optimality criterion to address consistent decision making under uncertainty over multiple filtrations.
- We derive an index which only involves a one dimensional (time-consistent) robust problem which is computationally tractable.
- Our model takes into account the desire to learn and uncertainty aversion.

Reference

- S.N. Cohen and T. Treetanthiploet, Gittins' theorem under uncertainty, arXiv:1907.05689.

Q & A

Sketch idea of the proof

Everything should be made as simple as possible, but not simpler—Albert Einstein

- Observe that $\gamma^{(m)}(t)$ is the minimum compensated reward which could encourage us to pay the cost from time t until the optimal stopping time τ^*.
- By the minimality of $\gamma^{(m)}(t)$, we must have zero total return under the optimal stopping, i.e.

$$\mathcal{E}^{(m)}\Big(\sum_{s=t+1}^{\tau^*}\beta^s(h^{(m)}(s)-\gamma^{(m)}(t))\Big|\mathcal{F}_t^{(m)}\Big)=0$$

Sketch idea of the proof

Everything should be made as simple as possible, but not simpler—Albert Einstein

- Observe that $\gamma^{(m)}(t)$ is the minimum compensated reward which could encourage us to pay the cost from time t until the optimal stopping time τ^*.
- By the minimality of $\gamma^{(m)}(t)$, we must have zero total return under the optimal stopping, i.e.

$$\mathcal{E}^{(m)}\Big(\sum_{s=t+1}^{\tau^*} \beta^s \big(h^{(m)}(s) - \gamma^{(m)}(t)\big) \Big| \mathcal{F}_t^{(m)} \Big) = 0$$

- In particular, at time τ^*, we need to increase the compensated reward to encourage further play.

Sketch idea of the proof

Everything should be made as simple as possible, but not simpler—Albert Einstein

- We increase the reward minimality by considering the reward process

$$\Gamma^{(m)}(t) = \max_{0 \le \theta \le t-1} \gamma^{(m)}(\theta).$$

- This minimal reward encourage us to pay the random cost $h^{(m)}(t)$ until the horizon $T^{(m)}$ with the return 0.

- We increase the reward minimality by considering the reward process

$$\Gamma^{(m)}(t) = \max_{0\le\theta\le t-1} \gamma^{(m)}(\theta).$$

- This minimal reward encourage us to pay the random cost $h^{(m)}(t)$ until the horizon $T^{(m)}$ with the return 0.
- Since this reward encourage us to continue at any point in time, for every $t \geq 0$,

$$\mathcal{E}^{(m)}\left(\sum_{s=t+1}^{T^{(m)}} \beta^s\left(h^{(m)}(s) - \Gamma^{(m)}(s)\right)\middle|\mathcal{F}_t^{(m)}\right) \le 0$$

with equality at $t = 0$.

Sketch idea of the proof

Everything should be made as simple as possible, but not simpler—Albert Einstein

- Now imagine that we are taking a break from the *m*-th arm to play another arm. Once we return to play the *m*-th arm we face a further rescaling of the discount factor. In particular, we have the discount factor $\alpha(s)\beta^s$ instead of β^s where α is decreasing.

Sketch idea of the proof

Everything should be made as simple as possible, but not simpler—Albert Einstein

- Now imagine that we are taking a break from the m-th arm to play another arm. Once we return to play the m-th arm we face a further rescaling of the discount factor. In particular, we have the discount factor $\alpha(s)\beta^s$ instead of β^s where α is decreasing.
- By the robust representation theorem, for any $\epsilon > 0$, show that there exists $\mathbb{Q} \in \mathcal{Q}^{(m)}$ such that

$$\mathbb{E}^{\mathbb{Q}}\Big[\sum_{s=1}^{T^{(m)}} \alpha(s)\beta^s\big(h^{(m)}(s) - \Gamma^{(m)}(s)\big)\Big] \geq -\epsilon$$

for every predictable decreasing process α in $[0,1]$.

Sketch idea of the proof

Everything should be made as simple as possible, but not simpler—Albert Einstein

- Now imagine that we are taking a break from the m-th arm to play another arm. Once we return to play the m-th arm we face a further rescaling of the discount factor. In particular, we have the discount factor $\alpha(s)\beta^s$ instead of β^s where α is decreasing.
- By the robust representation theorem, for any $\epsilon > 0$, show that there exists $\mathbb{Q} \in \mathcal{Q}^{(m)}$ such that

$$\mathbb{E}^{\mathbb{Q}}\Big[\sum_{s=1}^{T^{(m)}} \alpha(s)\beta^s(h^{(m)}(s) - \Gamma^{(m)}(s))\Big] \geq -\epsilon$$

for every predictable decreasing process α in $[0,1]$.
- In particular,

$$\mathfrak{E}\Big(\sum_n \beta^n(h^{(\rho_n)}(t_n^\rho) - \Gamma^{(\rho_n)}(t_n^\rho))\Big) \geq 0.$$

Sketch idea of the proof

Everything should be made as simple as possible, but not simpler—Albert Einstein

- If we take a break (i.e. leave the arm) only when

$$\mathcal{E}^{(m)}\left(\sum_{s=t+1}^{T^{(m)}} \beta^s \big(h^{(m)}(s) - \Gamma^{(m)}(s)\big)\Big|\mathcal{F}_t^{(m)}\right) = 0$$

i.e. when $\gamma^{(m)}$ reaches a new maximum, the total cost must be zero.

Sketch idea of the proof

Everything should be made as simple as possible, but not simpler—Albert Einstein

- If we take a break (i.e. leave the arm) only when

$$\mathcal{E}^{(m)}\Big(\sum_{s=t+1}^{T^{(m)}}\beta^s\big(h^{(m)}(s)-\Gamma^{(m)}(s)\big)\Big|\mathcal{F}_t^{(m)}\Big)=0$$

  i.e. when $\gamma^{(m)}$ reaches a new maximum, the total cost must be zero.
- In particular,

$$\mathfrak{E}\Big(\sum_n\beta^n\big(h^{(\rho_n^*)}(t_n^{\rho^*})-\Gamma^{(\rho_n^*)}(t_n^{\rho^*})\big)\Big)=0.$$

Sketch idea of the proof

Everything should be made as simple as possible, but not simpler—Albert Einstein

- If we take a break (i.e. leave the arm) only when

$$\mathcal{E}^{(m)}\Big(\sum_{s=t+1}^{T^{(m)}} \beta^s\big(h^{(m)}(s) - \Gamma^{(m)}(s)\big)\Big|\mathcal{F}_t^{(m)}\Big) = 0$$

  i.e. when $\gamma^{(m)}$ reaches a new maximum, the total cost must be zero.

- In particular,

$$\mathfrak{E}\Big(\sum_n \beta^n\big(h^{(\rho_n^*)}(t_n^{\rho^*}) - \Gamma^{(\rho_n^*)}(t_n^{\rho^*})\big)\Big) = 0.$$

- Finally, as $t \mapsto \Gamma^{(m)}(t)$ is increasing, ρ^* minimise $\sum_{n=1}^{N} \beta^n \Gamma^{(\rho_n^*)}(t_n^{\rho^*})$ for all N.