

Multi-Armed Bandits and the Gittins Index

Bobby Kleinberg

Cornell University



CS 6840, 28 April 2017

The Multi-Armed Bandit Problem



Multi-Armed Bandit Problem: A decision-maker (“[gambler](#)”) chooses one of n actions (“[arms](#)”) in each time step.

Chosen arm produces random payoff from unknown distribution.

Goal: [Maximize expected total payoff.](#)

Multi-Armed Bandits: An Abbreviated History

“The [MAB] problem was formulated during the war, and efforts to solve it so sapped the energies and minds of Allied scientists that the suggestion was made that the problem be dropped over Germany, as the ultimate instrument of intellectual sabotage.”

– P. Whittle



Multi-Armed Bandits: An Abbreviated History

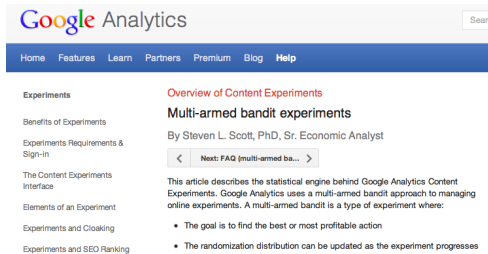
MAB algorithms proposed for [sequential clinical trials](#) in medicine: more ethical alternative to randomized clinical trials.



Multi-Armed Bandits: An Abbreviated History

On the Web, MAB algorithms used for, e.g.

- 1 Ad placement
- 2 Price experimentation
- 3 Crowdsourcing
- 4 Search



Google Analytics

Home Features Learn Partners Premium Blog Help

Experiments

Benefits of Experiments

Experiments Requirements & Sign-in

The Content Experiments Interface

Elements of an Experiment

Experiments and Cloaking

Experiments and SEO Ranking

Overview of Content Experiments

Multi-armed bandit experiments

By Steven L. Scott, PhD, Sr. Economic Analyst

Next: [FAQ \(multi-armed ba...](#)

This article describes the statistical engine behind Google Analytics Content Experiments. Google Analytics uses a multi-armed bandit approach to managing online experiments. A multi-armed bandit is a type of experiment where:

- The goal is to find the best or most profitable action
- The randomization distribution can be updated as the experiment progresses

The Bayesian Multi-Armed Bandit Problem (v1)

- Each arm has a *type* that determines its payoff distribution.
- Gambler has a prior distribution over types for each arm.
- Types are independent random variables.
- Objective: maximize expected discounted reward, $\sum_{t=0}^{\infty} \gamma^t r_t$.

The Bayesian Multi-Armed Bandit Problem (v1)

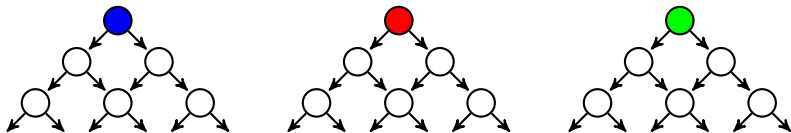
- Each arm has a *type* that determines its payoff distribution.
- Gambler has a prior distribution over types for each arm.
- Types are independent random variables.
- Objective: maximize expected discounted reward, $\sum_{t=0}^{\infty} \gamma^t r_t$.

After pulling arm i some number of times, gambler has a posterior distribution over types. Call this the *state* of arm i .

- State never changes except when pulled.
- When pulled, state evolution is a Markov chain. (Transition probabilities given by Bayes' Law.)
- Expected reward when pulled is governed by state.

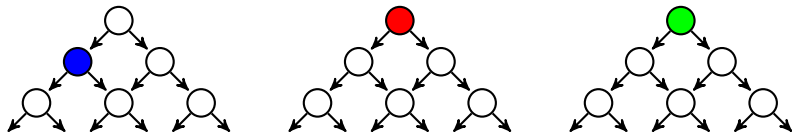
The Bayesian Multi-Armed Bandit Problem (v2)

- **Arm** = Markov chain and reward function $R : \{\text{States}\} \rightarrow \mathbb{R}$.
- **Policy** = function $\{\text{State-tuples}\} \rightarrow \{\text{arm to pull next}\}$.
- When pulling arm i at time t in state s_{it} , it yields reward $r_t = R(s_{it})$ and undergoes a state transition.
- **Objective**: maximize expected discounted reward, $\sum_{t=0}^{\infty} \gamma^t r_t$.



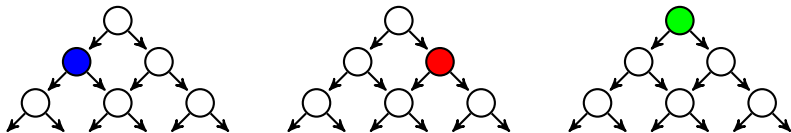
The Bayesian Multi-Armed Bandit Problem (v2)

- **Arm** = Markov chain and reward function $R : \{\text{States}\} \rightarrow \mathbb{R}$.
- **Policy** = function $\{\text{State-tuples}\} \rightarrow \{\text{arm to pull next}\}$.
- When pulling arm i at time t in state s_{it} , it yields reward $r_t = R(s_{it})$ and undergoes a state transition.
- **Objective**: maximize expected discounted reward, $\sum_{t=0}^{\infty} \gamma^t r_t$.



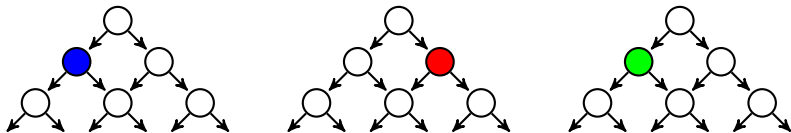
The Bayesian Multi-Armed Bandit Problem (v2)

- **Arm** = Markov chain and reward function $R : \{\text{States}\} \rightarrow \mathbb{R}$.
- **Policy** = function $\{\text{State-tuples}\} \rightarrow \{\text{arm to pull next}\}$.
- When pulling arm i at time t in state s_{it} , it yields reward $r_t = R(s_{it})$ and undergoes a state transition.
- **Objective**: maximize expected discounted reward, $\sum_{t=0}^{\infty} \gamma^t r_t$.



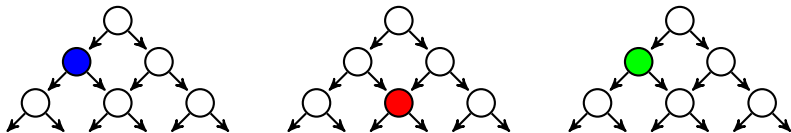
The Bayesian Multi-Armed Bandit Problem (v2)

- **Arm** = Markov chain and reward function $R : \{\text{States}\} \rightarrow \mathbb{R}$.
- **Policy** = function $\{\text{State-tuples}\} \rightarrow \{\text{arm to pull next}\}$.
- When pulling arm i at time t in state s_{it} , it yields reward $r_t = R(s_{it})$ and undergoes a state transition.
- **Objective**: maximize expected discounted reward, $\sum_{t=0}^{\infty} \gamma^t r_t$.



The Bayesian Multi-Armed Bandit Problem (v2)

- **Arm** = Markov chain and reward function $R : \{\text{States}\} \rightarrow \mathbb{R}$.
- **Policy** = function $\{\text{State-tuples}\} \rightarrow \{\text{arm to pull next}\}$.
- When pulling arm i at time t in state s_{it} , it yields reward $r_t = R(s_{it})$ and undergoes a state transition.
- **Objective**: maximize expected discounted reward, $\sum_{t=0}^{\infty} \gamma^t r_t$.



The Bayesian Multi-Armed Bandit Problem (v2)

In increasing order of generality, can assume each arm is ...

- 1 **Doob martingale of an i.i.d. sampling process**: state is conditional distribution of type given history of i.i.d. samples, R is conditional expected reward
- 2 **martingale**: arbitrary Markov chain, R satisfies **expected reward after state transition = current reward**
- 3 **arbitrary**: arbitrary Markov chain, arbitrary reward function.

This talk: mostly assume martingale arms.

The Gittins index

Consider a two-armed bandit problem where

- **arm 1** = Markov chain \mathcal{M} with starting state s , reward fcn. R
- **arm 2** yields reward ν deterministically when pulled.

Observation. Optimal policy pulls arm 1 until some stopping time τ , arm 2 forever after that if $\tau < \infty$.

The Gittins index

Consider a two-armed bandit problem where

- **arm 1** = Markov chain \mathcal{M} with starting state s , reward fcn. R
- **arm 2** yields reward ν deterministically when pulled.

Observation. Optimal policy pulls arm 1 until some stopping time τ , arm 2 forever after that if $\tau < \infty$.

Definition (Gittins index)

The Gittins index of state s in Markov chain \mathcal{M} is

$$\nu(s) = \sup\{\nu \mid \text{optimal policy pulls arm 1}\}.$$

Equivalently,

$$\nu(s) = \sup \left\{ \frac{\mathbb{E}[\sum_{t=0}^{\tau} \gamma^t r_t]}{\mathbb{E}[\sum_{t=0}^{\tau} \gamma^t]} \mid \tau \text{ a stopping time} \right\},$$

where r_0, r_1, \dots denotes the reward process of \mathcal{M} .

The Gittins Index Theorem

Theorem (Gittins Index Theorem)

For any multi-armed bandit problem with

- *finitely many arms*
- *reward functions taking values in a bounded interval $[-C, C]$*

a policy is optimal if and only if it always selects an arm with highest Gittins index.

Remark 1. Holds for Markov chains in general, not just martingales.

Remark 2. For martingale arms, $\nu(s) \geq R(s)$. The difference quantifies the value of foregoing short-term gains for future rewards.

Independence of Irrelevant Alternatives

One consequence of Gittins Index Theorem is an **independence of irrelevant alternatives** (IIA) property.

Corollary (IIA)

Consider a MAB problem with arm set U . If it is optimal to pull arm i in state-tuple \vec{s} , then it is also optimal to pull arm i in the MAB problem with any arm set V such that $\{i\} \subseteq V \subseteq U$.

In fact, the Gittins Index Theorem is **equivalent** to IIA.

Equivalence of Gittins Index Theorem and IIA

Consider state tuple (s_1, \dots, s_n) with Gittins indices ν_1, \dots, ν_n .

Let $\nu^{(1)} > \nu^{(2)}$ be the two largest values in the set $\{\nu_1, \dots, \nu_n\}$.

Add a deterministic arm with reward $\nu' \in (\nu^{(2)}, \nu^{(1)})$.

- 1 From this set of $n + 1$ arms, the optimal policy must choose one with index $\nu^{(1)}$. (Any other loses in a head-to-head comparison of two arms.)
- 2 From the original set of n arms, the optimal policy must choose one with index $\nu^{(1)}$.

Equivalence of Gittins Index Theorem and IIA

Consider state tuple (s_1, \dots, s_n) with Gittins indices ν_1, \dots, ν_n .

Let $\nu^{(1)} > \nu^{(2)}$ be the two largest values in the set $\{\nu_1, \dots, \nu_n\}$.

Add a deterministic arm with reward $\nu' \in (\nu^{(2)}, \nu^{(1)})$.

- 1 From this set of $n + 1$ arms, the optimal policy must choose one with index $\nu^{(1)}$. (Any other loses in a head-to-head comparison of two arms.)
- 2 From the original set of n arms, the optimal policy must choose one with index $\nu^{(1)}$.

Remark. IIA seems so much more natural than Gittins' Theorem, it's tempting to try proving IIA directly. Remarkably, no direct proof of IIA is known.

Proof of Gittins Index Theorem (Weber, 1992)

Consider a single-arm stopping game where the player can either

- 1 stop in any state s ,
- 2 pay ν , receive reward $R(s)$, observe next state transition.

A stopping rule is optimal if and only if it stops whenever $\nu(s) < \nu$ and keeps going whenever $\nu(s) > \nu$.

Proof of Gittins Index Theorem (Weber, 1992)

Consider a single-arm stopping game where the player can either

- 1 stop in any state s ,
- 2 pay ν , receive reward $R(s)$, observe next state transition.

A stopping rule is optimal if and only if it stops whenever $\nu(s) < \nu$ and keeps going whenever $\nu(s) > \nu$.

To encourage playing forever, whenever we arrive at state s such that $\nu(s) < \nu$ we could lower the charge from ν to $\nu(s)$.

Definition (Prevailing charge process)

If s_0, s_1, \dots denotes a state sequence sampled from Markov chain \mathcal{M} the **prevailing charge process** is the sequence $\kappa(0), \kappa(1), \dots$ defined by

$$\kappa(t) = \min\{\nu(s_0), \nu(s_1), \dots, \nu(s_t)\}.$$

Proof of Gittins Index Theorem (Weber, 1992)

In the **prevailing charge game** for Markov chain \mathcal{M} , the charge for playing at time t is $\kappa(t)$.

Lemma

In the prevailing charge game, the expected net payoff of any stopping rule is non-positive. It equals zero iff the rule never stops in a state whose Gittins index exceeds the prevailing charge.

Proof idea. Break time into intervals on which $\kappa(t)$ is constant.

Proof of Gittins Index Theorem (Weber, 1992)

In the **prevailing charge game** for Markov chain \mathcal{M} , the charge for playing at time t is $\kappa(t)$.

Lemma

In the prevailing charge game, the expected net payoff of any stopping rule is non-positive. It equals zero iff the rule never stops in a state whose Gittins index exceeds the prevailing charge.

Proof idea. Break time into intervals on which $\kappa(t)$ is constant.

Remark. Lemma still holds if there is also a “pausing” option at every time. A strategy breaks even if and only if it never stops or pauses except when $\nu(s_t) = \kappa(t)$.

Proof of Gittins Index Theorem (Weber, 1992)

Proof of Gittins Index Theorem. Suppose π is the Gittins index policy and σ is any other policy.

Couple the executions of π, σ by assuming each arm goes through the same state sequence. (So π, σ differ only in how they interleave the state transitions.)

Theorem will follow from

$$\begin{aligned}\mathbb{E}[\text{Reward}(\pi)] &= \mathbb{E}[\text{PrevChg}(\pi)] \\ \mathbb{E}[\text{Reward}(\sigma)] &\leq \mathbb{E}[\text{PrevChg}(\sigma)] \\ \mathbb{E}[\text{PrevChg}(\sigma)] &\leq \mathbb{E}[\text{PrevChg}(\pi)].\end{aligned}$$

First two lines arise from our lemma.

Third line arises from our coupling.

Proof of Gittins Index Theorem (Weber, 1992)

Prevailing charge sequences:

Arm 1: 5,3,3,1,1,1,...

Arm 2: 4,4,4,2,2,2,...

Arm 3: 6,3,1,1,1,1,...

Proof of Gittins Index Theorem (Weber, 1992)

Prevailing charge sequences:

Arm 1: **5,3,3**,1,1,1,...

Arm 2: **4,4,4,2,2,2**,...

Arm 3: **6,3**,1,1,1,1,...

π : **6,5,4,4,4,3,3,3,2,2,2**,...

Proof of Gittins Index Theorem (Weber, 1992)

Prevailing charge sequences:

Arm 1: **5,3,3,1**,1,1,...

Arm 2: **4,4,4,2,2,2**,...

Arm 3: **6,3**,1,1,1,1,...

π : **6,5,4,4,4,3,3,3,2,2,2**,...

σ : **5,4,6,3,4,3,3,4,2,1,2**,...

Proof of Gittins Index Theorem (Weber, 1992)

Prevailing charge sequences:

Arm 1: **5,3,3,1**,1,1,...

π : **6,5,4,4,4,3,3,3,2,2,2**,...

Arm 2: **4,4,4,2,2,2**,...

σ : **5,4,6,3,4,3,3,4,2,1,2**,...

Arm 3: **6,3**,1,1,1,1,...

$$\text{PrevChg}(\pi) = 6 + 5\gamma + 4\gamma^2 + 4\gamma^3 + 4\gamma^4 + 3\gamma^5 + \dots$$

$$\text{PrevChg}(\sigma) = 5 + 4\gamma + 6\gamma^2 + 3\gamma^3 + 4\gamma^4 + 3\gamma^5 + \dots$$

Policy π sorts the prevailing charges in decreasing order.

This maximizes the discounted sum, QED.

Gittins Index Calculation: “Collapsing” Arms

A *collapsing arm* is one of two types: G (good) or B (bad).

G always yields reward M , B always yields reward 0.

$\Pr(G) = p, \Pr(B) = 1 - p$.

What is its Gittins index? If arm 1 is collapsing and arm 2 is deterministic with reward ν , compare:

- 1 π : Play arm 1 once, play the better arm thenceforward.

$$V(\pi) = p \left(\frac{M}{1-\gamma} \right) + (1-p) \left(\frac{\gamma}{1-\gamma} \right) \nu$$

- 2 σ : Play arm 2 forever. $V(\sigma) = \left(\frac{1}{1-\gamma} \right) \nu$

Equating the two values, $\nu = \frac{pM}{1-(1-p)\gamma}$.

Gittins Index Versus Expected Reward

For martingale arms, $\nu(s) \geq R(s)$. The difference quantifies the value of foregoing short-term gains for future rewards.

What are the possible values of $\nu(s)$ for martingales?

Initial state of a collapsing arm has $R(s) = pM$, $\nu(s) = \frac{pM}{1-(1-p)\gamma}$, which shows that $\nu(s)$ can take any value in $[R(s), \frac{R(s)}{1-\gamma}]$.

No other ratio is possible.

If arm 2 yields reward $\nu \geq \frac{R(s)}{1-\gamma}$ and π plays arm 1 at time 0, then $\mathbb{E}[r_t]$ is $R(s)$ at $t = 0$ and less than $R(s) + \nu$ afterward, so

$$V(\pi) < \frac{R(s)}{1-\gamma} + \frac{\gamma\nu}{1-\gamma} \leq \nu + \frac{\gamma\nu}{1-\gamma} = V(\sigma),$$

where σ is the policy that always plays arm 2.

Multi-Armed Bandit Problem

- Fundamental abstraction of sequential learning with “explore vs. exploit” dilemmas.
- Modeled by n -tuple of Markov chains with rewards on states.
- Arms undergo state transitions only when pulled.
- Objective: maximize geometric (rate γ) discounted reward.

Gittins Index Policy

- A reduction from multi-armed bandits to two-armed bandits.
- Gittins index $\nu(s)$ defined so that optimal policy is indifferent between arm in state s or deterministic arm with reward $\nu(s)$.
- Optimal policy: always pull arm with highest Gittins index.

Fragility of the Gittins Index Theorem

Warning! The Gittins Index Theorem is beautiful but non-robust.

Vary any assumption, and you get a problem to be deployed against enemy scientists in the present day!

Fragility of the Gittins Index Theorem

Warning! The Gittins Index Theorem is beautiful but non-robust.

Vary any assumption, and you get a problem to be deployed against enemy scientists in the present day!

Examples

- 1 non-geometric discounting, e.g. fixed time horizon.
- 2 arms with correlated priors
- 3 actions affecting more than one arm at a time
- 4 payoffs depending on state of 2 or more arms
- 5 delayed feedback
- 6 “restless” arms that change state without being pulled
- 7 non-additive objectives (e.g. risk-aversion)
- 8 switching costs