Stat 260/CS 294-102. Learning in Sequential Decision Problems.

Peter Bartlett

1. Gittins Index:
   - Discounted, Bayesian (hence Markov arms).
   - Reduces to stopping problem for each arm.
   - Interpretation as (scaled) equivalent lump sum.
   - Computation.

Gittins index for Bayesian bandits

At time t, arm j gives $X_{j,t}$ $\sim$Bernoulli(p_j) reward. Aim to choose sequence of arms $I_1, I_2, \ldots,$ so as to maximize total expected discounted reward:

$$\mathbb{E} \sum_{t=0}^{\infty} \gamma^t X_{I_t,t},$$

where $0 < \gamma < 1$ is a discount factor that ensures limits exist.

Gittins index for Bayesian bandits

Assume $\mathrm{Beta}(\alpha, \beta)$ prior on p_j:

$$\pi(p|\alpha,\beta) = \frac{p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha,\beta)},$$

where B is the *beta function*:

$$B(\alpha,\beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}\,dx = \frac{(\alpha-1)!(\beta-1)!}{(\alpha+\beta-1)!}$$

for positive integers α, β. This is a conjugate prior for the binomial distribution: the posterior distribution of p given k successes out of n is

$$\begin{aligned} P(p|k,\alpha,\beta) &\propto P(k|p)\pi(p|\alpha,\beta) \\ &\propto p^k(1-p)^{n-k}p^{\alpha-1}(1-p)^{\beta-1}, \end{aligned}$$

which is a $\mathrm{Beta}(\alpha+k, \beta+n-k)$ distribution.

Gittins index for Bayesian bandits

Assume p_i independent. Then if arm j has been pulled n times with k successes, we can think of $s_{j,n} = (\alpha + k, \beta + n - k)$ as the state of the arm, and we know how the state evolves when the arm is pulled:

$$P\left(s_{j,n+1} = s_{j,n} + (1,0) | s_{j,n}\right) = P(X_{j,n} = 1 | s_{j,n}).$$

And we also know the reward distribution

$$P\left[X_{j,n} = 1 | s_{j,n}\right].$$

This is just the mean of the Beta posterior:

$$\mathbb{E}\left[X_{j,n} | s_{j,n} = (\alpha, \beta)\right] = \mathbb{E}\left[p_j | s_{j,n} = (\alpha, \beta)\right] = \frac{\alpha}{\alpha + \beta}.$$

Gittins index for Bayesian bandits

From now on, we'll assume:

1. that the state $s_j(t)$ of arm j is unchanged if we do not choose that arm,
2. that we choose a single arm, its state evolves as a known Markov chain when we choose it, and our actions do not affect that evolution.
3. that given the state $s_{I_t}(t)$ of the arm that is played, the reward $X_{I_t,t}$ is conditionally independent of the past and of the other arms' states: $X_{I_t,t} = R_{I_t}(s_{I_t}(t))$.

Choose a policy π to maximize the total expected discounted reward,

$$V_\pi(s) = \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t R_{I_t}(s_{I_t}(t)) \middle| s(0) = s\right].$$

We'll call a problem of this kind a *discounted Markov bandit problem*.

Gittins index for Bayesian bandits

We might be tempted to use dynamic programming to find the optimal value:

$$V(s) = \max_j \left\{ \mathbb{E}R_j(s_j) + \gamma \sum_{s'_j} P_j(s'_j|s_j)V(s') \right\}.$$

where $s = (s_1, \ldots, s_k)$ and $s' = (s_1, \ldots, s_{j-1}, s'_j, s_{j+1}, \ldots, s_k)$.

But the state space has size exponential in k!

Markov bandit problems are easier...

Gittins index: some intuition

If we knew the sequence of rewards, we could balance the size and timing of the rewards to maximize $\sum_t \gamma^t R_{I_t}$.

Example:

$$\begin{array}{ccccccccccccc} 8 & 8 & 8 & 7 & 7 & 7 & 6 & 6 & 5 & 3 & 3 & 3 & \cdots \\ 7 & 6 & 6 & 6 & 6 & 5 & 4 & 4 & 4 & 4 & 2 & 2 & \cdots \end{array}$$

Example:

$$\begin{array}{ccccc} 6 & 7 & 7 & 7 & \cdots \\ 7 & 6 & 6 & 6 & \cdots \end{array}$$

Gittins index: some intuition

Decreasing reward sequences are easy. How could we make the rewards decreasing?

> If, once we've chosen an arm, we keep choosing it as long as it looks at least as good as it looked when we started playing it, then how good it looks is non-increasing.

(Of course, "how good it looks" at the start will depend on how long we anticipate playing it.)

Let's consider a simpler problem, where we need to choose the order of the arms (and we never return to an arm).

Gittins index: some intuition

Consider the related (simpler) problem of scheduling jobs on a machine: Job i takes time t_i and, on completion, gives reward r_i.

How should we order them, so as to maximize total discounted reward?

1 then 2:

$$r_1\gamma^{t_1} + r_2\gamma^{t_1+t_2} \text{ vs } r_2\gamma^{t_2} + r_1\gamma^{t_1+t_2}$$

We should schedule 1 before 2 if

$$\frac{\gamma^{t_1}}{1-\gamma^{t_1}} r_1 > \frac{\gamma^{t_2}}{1-\gamma^{t_2}} r_2.$$

So we can compute this index for each job, and schedule them in decreasing order of the indices.

Gittins index

Gittins' work reduced discounted Markov bandit problems to stopping problems, and used a swapping argument to show the optimality of a 'dynamic allocation index' (Gittins index).

Theorem: [Gittins Index Theorem] For any discounted Markov bandit problem, define the "Gittins index":

$$G_j(s_j) = \sup\left\{\alpha : \sup_\tau \mathbb{E}\left[\sum_{t=0}^{\tau-1} \gamma^t \left(R_j(s_j(t)) - \alpha\right) \middle| s_j(0) = s_j\right] \geq 0\right\},$$

where $\tau \geq 1$ is a *stopping time*. Then there is an optimal policy that, at time t, chooses

$$I_t \in \arg\max_j G_j(s_j(t)).$$

Some properties of the Bernoulli case

For s successes and f failures:

Success helps: $G((s, f+1)) < G((s, f)) < G((s+1, f))$

For $s + f \to \infty$, $G((s, f)) \to \dfrac{s}{s+f}$

Failure hurts for large γ:

$\forall k > 0,\ \exists \gamma^*,\ \forall \gamma > \gamma^*,\ G((s+k, f+1)) < G((s, f))$

Gittins index proof

What does the index mean?

$$G_j(s_j) = \sup \left\{ \alpha : \sup_{\tau} \mathbb{E} \left[\sum_{t=0}^{\tau-1} \gamma^t \left(R_j(s_j(t)) - \alpha \right) \middle| s_j(0) = s_j \right] \geq 0 \right\},$$

Fix an arm. Think of α as a fixed tax. Consider a stopping game:

- Suppose the tax is α.
- At time t, if I haven't already stopped (that is, $\tau > t$), I can choose to keep playing, pay the tax α and receive the reward $R_j(s_j(t))$.
- τ is when I choose to stop. (Can depend only on the *state*.)

Suppose I keep playing if the tax is at or below my subsequent expected discounted reward, and stop as soon as the tax becomes excessive.

Gittins index proof

Crucial properties:

1. Expected total discounted profit is zero
   (because of the way the tax is set for the starting state).
2. The value of α can only decrease as this game progresses.
   (If the fair tax level increases above α, I get to continue playing and make a profit. It is only when it drops below α that I stop playing.)

Gittins index proof

Multiple arms:

Each arm offers a fair game, provided that I play it optimally (continue to play it while the tax is fair or favorable).
In that case, the expected total discounted profit is still zero.
Expected discounted reward = expected discounted tax paid.

For each arm, the sequence of taxes is:

1. non-increasing,
2. random,
3. independent of the policy.

Non-increasing ⇒ there is a unique interleaving of these sequences into a single non-increasing sequence of taxes, and this corresponds to the largest total discounted tax paid (because of the discount factor).

Gittins index proof

The strategy that plays the arm with the highest tax (until the optimal stopping time) is equivalent to the strategy that plays the arm with the highest Gittins index G_j (if G_j was the highest and it increases, then with optimal stopping, we would continue to play j). And this strategy will pay the largest total discounted tax, that is, will have the maximal total discounted expected reward.

Gittins index: A retirement interpretation

$$G_j(s_j) = \sup_{\tau} \frac{\mathbb{E}\left[\sum_{t=0}^{\tau-1} \gamma^t R_j(s_j(t)) \middle| \, s_j(0) = s_j\right]}{\mathbb{E}\left[\sum_{t=0}^{\tau-1} \gamma^t \middle| \, s_j(0) = s_j\right]},$$

which is the maximal ratio (under optimal choice of stopping time τ) of expected discounted reward to expected discounted time.

If we define $L = G_j(s_j)/(1-\gamma)$, then since
$(1-\gamma)\sum_{t=0}^{\tau-1} \gamma^t L = (1-\gamma^\tau)L$,

$$L = \mathbb{E}\left[\sum_{t=0}^{\tau-1} \gamma^t R_j(s_j(t)) + \gamma^\tau L \middle| \, s_j(0) = s_j\right].$$

So this is the value L for which we would consider receiving the lump sum L now or receiving L after some optimal number of further rewards to be equally good alternatives.

Computing the Gittins index

How do we calculate G_j?

Offline Calculate the table of values of the index for each state.

Online Calculate the value of the index for the current state, and the corresponding stopping time (equivalently, stopping set) for the current state. This is convenient when the state space is large.

Notice that the optimal stopping problem is a problem of controlling a Markov decision process. If the state space is not too large, we could use value iteration (iterate the Bellman optimality equations), or solve the corresponding linear program. Chen and Katehakis (1986) showed that this can be extended to include the optimization of the value α as part of the LP. Another approach, due to Varaiya, Walrand and Buyukkoc, involves properties of the stopping time: for a state s, the stopping set is all states with Gittins index lower than s.

Computing the Gittins index: Offline

Largest Remaining Index Algorithm:

1. Find state s_1 with largest Gittins index, $G(s_1) = R(s_1)$.
2. Given states $s_1, \ldots, s_{k-1}$ with largest Gittins indices, find s_k as follows:
   (a) Define the continuation set $C(s_k) = \{s_1, \ldots, s_{k-1}\}$.
   (b) Define the continuation transition matrix by $P^{(k)}_{s',s} = P_{s,s'} 1[s \in C(s_k)]$.
   (c) Compute the values and durations
   $$V^{(k)} = (I - \gamma P^{(k)})^{-1} R, \qquad d^{(k)} = (I - \gamma P^{(k)})^{-1} 1.$$
   (d) Set $s_k = \arg\max_{s \in C(s_k)} V_s^k / d_s^{(k)}$, with $G(s_k) = V_s^k / d_s^{(k)}$.

Computing the Gittins index: Online

Recall that, for $L = G_j(s_j)/(1-\gamma)$,

$$L = \mathbb{E}\left[\sum_{t=0}^{\tau-1} \gamma^t R_j(s_j(t)) + \gamma^\tau L \middle| s_j(0) = s_j \right].$$

For a fixed L, the optimal stopping time will achieve value from each start state given by the unique vector v satisfying $v = \max\{R + \gamma P v, L1\}$.

Suppose we want to compute $G(s)$ for a single state s. Observe that, if L is the correct retirement payout for s, then either retiring or restarting in s will lead to the same subsequent total discounted expected reward. So if we allow, in each state, to restart in state s, the solution to the fixed point equation $v = \max\{R + \gamma P v, 1R(s) + \gamma P_s' v\}$ gives $G(s) = (1-\gamma)v_s$. And we could solve this fixed point equation by formulating it as a linear program, for example.