The Gittins Strategy: Choosing from Multiple Markov Chains

Seth Harris

March 16, 2011

1 Introduction: Markov systems and games

Suppose you are playing a game of chance, such as taking a random walk to reach a target. You can often choose among multiple strategies to reach the same goal, and often it is in your best interest to continually change your strategy depending on how your game progresses.

For example, consider the following random walk on a number line with vertices from 0 to 5 (see Figure 1 below). There is a prize at vertex 3, and we have two starting tokens at vertices 2 and 5. Elementary proababillty theory shows that if we start at vertex 2, we can expect to reach the prize in $3^2 - 2^2 = 5$ moves, whereas if we start at vertex 5, we can expect to reach the prize in $2^2 = 4$ moves. However, we shall show later that the optimal strategy is to move the token at vertex 2 once, and if it happens to go to the left, switch to the token at vertex 5 until the end. With this strategy we can expect to hit the prize in $1 + \frac{1}{2}(2^2) = 3$ moves.

Dumitriu, Tetali, and Winkler [1] introduce a quantity called the *grade* of the Markov system, which in some way measures the desirablility of playing that system. It is a small variation on the more famous Gittins index (see [2]), another metric of the reward of a stochastic process. The Gittins index differs from our grade in that its Markov systems have no target, and have a probability of terminating at every stage and an associated discount β. We will use the name "Gittins strategy" for the optimal strategy of choosing between Markov games.

We will be dealing with Markov systems $\mathcal{S} = \langle V, P, C, s, t \rangle$, with V a set of states, P a transition matrix between states, C a vector of costs for moving to each state, a source

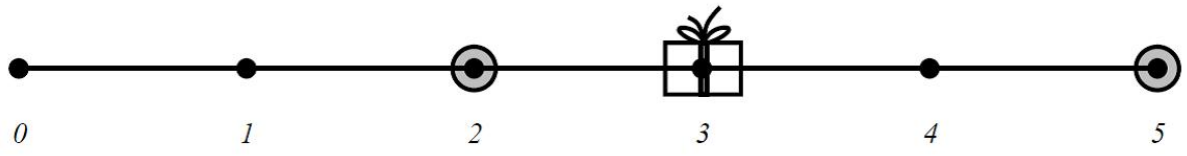

Figure 1: We can get to the gift using two possible tokens.

state s and a target state t. Our goal is to minimize the total cost of a random walk from the source s to the target t. When we combine several Markov processes, we can view it as a game $\mathcal{G}$ in which in which we choose a strategy (i.e. choose which token to move) for reaching one of the targets. $\mathcal{S}_1 \circ \mathcal{S}_2 \circ \ldots \circ \mathcal{S}_k$ is a multitoken Markov game: we choose one of k tokens, follow the associated Markov chain, and pay the associated cost. On the next move, we can choose any other token to move; upon reaching any of the targets, we stop. For any game $\mathcal{G}$, $\mathbf{E}[\mathcal{G}]$ is defined to be the minimum expected cost over *all* possible strategies; a strategy which achieves this minimum is said to be *optimal*. The *join* of multiple games, $\mathcal{G} = \mathcal{G}_1 \square \mathcal{G}_2 \square \ldots \square \mathcal{G}_n$, is defined similarly: choose to play one of the $\mathcal{G}_i$'s, switch games at any time, and the join ends when any of its games ends. A player is *indifferent* among a set of moves if each move is part of an optimal strategy. One particularly simple game that we will use in our proof is the *terminator* game $\mathcal{T}_g = \langle \{s, t\}, P, g, s, t \rangle$. Here $p_{s,t} = 1$; in other words, we just pay a cost of g to make a single move.

2 The grade and the Gittins strategy

The key metric for comparing multiple Markov systems is the grade.

Definition 2.1. The *grade* $\gamma(\mathcal{S})$ of a Markov system $\mathcal{S}$ is the unique value of g such that a player is indifferent in $\mathcal{G}_g = \mathcal{S} \circ \mathcal{T}_g$.

So there is a unique cost, g, such that that a player is indifferent between simply paying g dollars at the outset, or playing the game $\mathcal{S}$ and paying all its associated costs. It can be shown that for any strategy σ for $\mathcal{G}_g$, $E[\sigma]$ is linear in g; therefore, the minimum expected cost over all strategies, $\mathbf{E}[\mathcal{G}_g]$, is piecewise linear in g. On the following page (Figure 2) is a typical graph of $\mathbf{E}[\mathcal{G}_g]$ versus g. When g is too high, one always chooses to play $\mathcal{S}$; hence the graph is horizontal at $E[\mathcal{S}]$. When g is too low, one always chooses to pay g up front; hence the graph has slope 1. This leftmost piece of the graph has maximum value $\gamma(\mathcal{S})$, the grade, representing indifference of playing $\mathcal{S}$ or paying g. For intermediate values of g, one starts by playing $\mathcal{S}$ but may switch depending on how the game progresses.

We can also talk about the grade $\gamma_u(\mathcal{S})$, the grade of the game starting at vertex u rather than the source. Thus $\gamma_s(\mathcal{S}) = \gamma(\mathcal{S})$. For a given g, the optimal strategies for $\mathcal{G}_g$ are easy to characterize:

Proposition 2.2. *A strategy for $\mathcal{G}_g = \mathcal{S} \circ \mathcal{T}_g$ is optimal if and only if it chooses $\mathcal{S}$ whenever we are in a state u with $\gamma_u < g$ and chooses $\mathcal{T}_g$ whenever $\gamma_u > g$.*

Our major result is the following:

Theorem 2.3. *A strategy for $\mathcal{G} = \langle \mathcal{S}_1 \circ \ldots \circ \mathcal{S}_n \rangle$ is optimal if and only if it always plays a system whose current grade is minimal.*

This strategy is known as the Gittins strategy, named for the Gittins index described in the introduction. We will use a series of easy lemmas to prove 2.3. We start by reformulating each original game as a "reward game" $\mathcal{S}_i(g)$, where we pay and play as in $\mathcal{S}_i$ and may quit at any time, but there is a reward of g at the target.

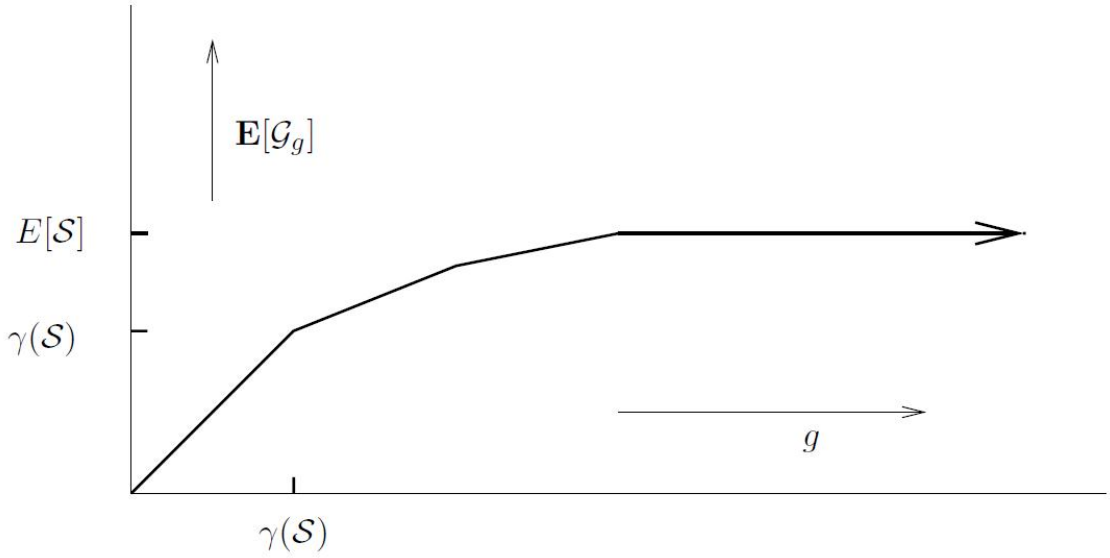

Figure 2: The expected cost of $\mathcal{G}_g$ is piecewise linear in g.

Lemma 2.4. *Let $g = \gamma(\mathcal{S}_i)$. Then $\mathcal{S}_i(g)$ is a fair game; that is, $\boldsymbol{E}[\mathcal{S}_i(\gamma(\mathcal{S}_i))] = 0$. Moreover, a strategy is optimal if and only if the player quits whenever $\gamma_u > g$ and plays on when $\gamma_u < g$.*

This is fairly easy to see: it is essentially a translation of our terminator game, only instead of having a cost of $\gamma(\mathcal{S}_i)$ upon quitting and no reward at the target, we have no cost upon quitting and a reward of $\gamma(\mathcal{S}_i)$ at the end. The optimal strategies are clearly equal, and by Proposition 2.2, the strategy is to quit when $\gamma_u > g$ and play on when $\gamma_u < g$. Since one optimal strategy is to quit right away (because then one is indifferent), the expectation of the game is 0; hence the game is fair.

We now slightly change our game to a "teasing" reward game $\mathcal{S}_i'$: whenever you reach a state u with $\gamma_u > g$, your reward at the end is boosted up to γ_u.

Lemma 2.5. *$\mathcal{S}_i'$ is also a fair game.*

As the original reward game $\mathcal{S}_i(\gamma(\mathcal{S}_i))$ is fair, and since the additional reward increases precisely when the expected cost increases, the new game is also fair. Now we analyze the join of these teasing games:

Lemma 2.6. *$\mathcal{G}' = \mathcal{S}_1' \square \ldots \square \mathcal{S}_n'$ is also fair.*

$\mathcal{G}'$ clearly cannot be worse than fair, since one can just choose to play $\mathcal{S}_1$ at every move. But it cannot be better than say $\mathcal{S}_1$, as it can only cost more before the target is reached.

To show that the Gittins strategy is optimal for our original game $\mathcal{G}$, we first note that it is optimal for $\mathcal{G}'$:

Lemma 2.7. *The Gittins strategy Γ is optimal for $\mathcal{G}'$.*

Lemmas 2.4 and 2.5 together imply that Γ plays each individual $\mathcal{S}_i'$ optimally, since it chooses to stay in $\mathcal{S}_i'$ if and only if the current grade is below the reward. So Γ will also play $\mathcal{G}'$ optimally.

Lemma 2.8. *Of all nonquitting strategies for $\mathcal{G}'$, the Gittins strategy Γ reaps the smallest expected reward in the end. Moreover, Γ is the* only *nonquitting strategy with this property.*

Thus without the move-costs, Γ would actually be the worst strategy for $\mathcal{G}'$. The intuition behind this lemma is that the higher rewards always correspond to the grade going up. Since we're using the one strategy which deliberately always chooses the *lowest possible* grade, it must also be the unique strategy with the lowest possible expected reward.

We are now ready to prove that Γ is an optimal strategy for our original game $\mathcal{G} = \mathcal{S}_1 \circ \ldots \circ \mathcal{S}_n$, and in fact the unique one:

Proof of Theorem 2.3. Let us recall that in $\mathcal{G}$, there are no rewards at all, and that an optimal strategy for $\mathcal{G}$ will simply choose among $\mathcal{S}_1, \ldots, \mathcal{S}_n$ in a way that minimizes the total cost. For any nonquitting strategy Δ for $\mathcal{G}'$, let $C(\Delta)$ be its expected cost and let $R(\Delta)$ be its expected reward. Since $\mathcal{G}'$ is fair, $E[\Delta] = R(\Delta) - C(\Delta) \leq 0$. But by Lemma 2.7, we have $E[\Gamma] = 0$ in $\mathcal{G}'$; thus $C(\Gamma) = R(\Gamma) \leq R(\Delta) \leq C(\Delta)$, the middle inequality following from Lemma 2.8. But $C(\Gamma) \leq C(\Delta)$ precisely shows that Γ is an optimal strategy for our original game $\mathcal{G}$.

Moreover, if Δ is another optimal strategy, then the above inequalities must be equalities. So Δ also reaps the minimum reward for $\mathcal{G}'$, and so by Lemma 2.8, Δ is also Gittins strategy. So we have the converse.

□

3 Computing the grade

The grades of all states of a Markov system can be computed in polynomial time, specifically $\mathcal{O}(n^5)$. The following theorem will be crucial in defining our algorithm:

Let $\{s, t\} \subseteq U \subseteq V$. Define a new game $\mathcal{S} \restriction U$ as follows: If we are currently in U and try to travel outside U, we restart at s. Formally this means that the transition matrix is equal to that of $\mathcal{S}$ for states outside of U, but for each $u \in U$, the transition probability $p_{u,s}$ is increased by $\Sigma_{v \in V \setminus U} p_{u,v}$.

Theorem 3.1. *With $\mathcal{S}$ and U as above, $\gamma_s(\mathcal{S}) \leq E_s[\mathcal{S} \restriction U]$, with equality if U contains all states of grade lower than γ_s and no states of grade higher than γ_s.*

Proof. Imagine a reward game that pays γ_s when we reach our target, and consider the strategy σ that chooses to quit the game whenever we restart. Since the reward game is fair, it has nonpositive expectation. If we play of series of these games until we reach our target, the resulting game will still have nonpositive expectation, the total reward will be $\gamma_s(\mathcal{S})$, and the total expected cost will be $E_s[\mathcal{S} \restriction U]$; therefore, we have our inequality. On the other hand, if U is as described, we know by Lemma 2.4 that σ is optimal (in both situations we quit if the grade is greater than the reward), and so we have an equality. □

This idea of restricting our game to a subset of states will enable us to compute the grade in polynomial time.

Theorem 3.2. *For a Markov system $\mathcal{S}$, the following algorithm will compute $\gamma(\mathcal{S})$.*

Proof. This will hopefully be a clear sketch of the algorithm; for a complete pseudocode see [1]. For any given set of states U, let $N(U)$, the neighborhood of U, be the set of states that can reach U in one move; i.e., $N(U) = \{v \in V \mid p_{v,u} > 0 \text{ for some } u \in U\}$.

1. Begin with $U = \{t\}$; clearly $\gamma_t = 0$.
2. For each state v in $N(U)$, compute $E_v[\mathcal{S}']$, where $\mathcal{S}' = \mathcal{S} \restriction (U \cup \{v\})$.
3. Compare all values of $E_v[\mathcal{S}']$ among states $v \in N(U)$. Choose the v with the least such value and add it to U. Declare this value to be the grade $\gamma_v(\mathcal{S})$.
4. Repeat steps 1-3 until $U = V$.

Why does our "declaring" $E_v[\mathcal{S}']$ to be $\gamma_v(\mathcal{S})$ work? By Theorem 3.1, as long as U contains all states of grade lower than γ_v and no states higher than γ_v, we have $\gamma_v = E_v[\mathcal{S}']$. This is clearly true for our U as long as we're sure of one thing: that among all states not in U, the state of lowest grade is in $N(U)$. Observe that as long as we can go from the source to the target in finite time, there will be a path from s to t of decreasing grade. Thus if the smallest-graded state not in U were outside $N(U)$, there would be a decreasing-graded path into U, which would pass through a smaller-graded state in $N(U)$, a contradiction. Therefore, our algorithm accurately computes the grades of all states in V.

□

To analyze the computing time, suppose we are in step i of the algorithm out of a total of n steps. The most costly part of the algorithm is finding $E_v[\mathcal{S}']$, which involves solving an $(i+1) \times (i+1)$ system of equations, requiring $\mathcal{O}(i^3)$ steps. In step i, we have to compute this for every state in $N(U)$, and $|N(U)| = \mathcal{O}(n-i)$. So the total computation time will be $\mathcal{O}(\Sigma_{i=1}^{n}(n-i)i^3) = \mathcal{O}(n^5)$.

To illustrate this algorithm, let us return to the random walk on vertices 0 to 5 in Figure 1. While we will not compute every possible grade, we will indeed prove that the optimal strategy is to start at the token at 2, and if that fails, play the token at 5. I will call the two Markov systems $\mathcal{S}$ and $\mathcal{T}$, with $\mathcal{S}$ representing moving the token on 2 and $\mathcal{T}$ representing moving the token on 5; thus we are playing the game $\mathcal{S} \circ \mathcal{T}$. Letting our algorithm guide us, we start by computing the grades at points that are neighbors to 3: computing $\gamma_2(\mathcal{S})$ and $\gamma_4(\mathcal{T})$. (Computing $\gamma_i(\mathcal{S})$ for $i > 3$ is irrelevant, since our token will reach our target before ever going there; similarly computing $\gamma_j(\mathcal{T})$ for $j < 3$ is irrelevant.)

To find the grade $\gamma_2(\mathcal{S})$, we consider the game restricted to $U = \{2, 3\}$, so that whenever we go to the left to vertex 1, we effectively have a loop bringing us back to 2. Our game thus is a Bernoulli process (success moving to the right, failure looping back to 2) which has expected time 2. So $\gamma_2(\mathcal{S}) = 2$; similarly, $\gamma_4(\mathcal{T}) = 2$.

Next comes computing the grade two vertices from the target. For $\gamma_5(\mathcal{T})$, our restricted neighborhood U is $\{2,3,4,5\}$, so it's really not a restricted game at all, and the expected time of the game is 4, just like it is normally. Now for $\gamma_1(\mathcal{S})$, our restricted neighborhood U is $\{1,2,3,4\}$. So effectively we're looking for the hitting time on the line from 1 to 3, plus a loop at 1. It is easy to see that this hitting time must be strictly greater than 4.

The above analysis suffices to deduce the Gittins strategy for $\mathcal{S} \circ \mathcal{T}$. Our initial grades are $\gamma_2(\mathcal{S}) = 2$ and $\gamma_5(\mathcal{T}) = 4$, so we initially choose to move the token on 2. If we happen to move to the left, then compare our grades: $\gamma_1(\mathcal{S}) > 4$, $\gamma_5(\mathcal{T}) = 4$, and $\gamma_4(\mathcal{T}) = 2$. Therefore, we always will choose to play the game $\mathcal{T}$, since it will always be the game of minimal grade. Thus we have shown that this strategy is the optimal one in the game $\mathcal{S} \circ \mathcal{T}$, and indeed the unique one, since we were never indifferent between the two games at any point.

4 A small note on infinite random walks

Thus far, we have assumed that our Markov systems are finite. However, the grade is defined for infinite Markov systems as long as they are locally finite; that is, as long as each state can only move to and from finitely many other states:

Theorem 4.1. *Let $\mathcal{S}$ be an infinite, locally finite Markov system with a designated target state t. Then $\gamma_u < \infty$ for all $u \in V$. Moreover, if we fix $u \in \mathcal{S}$, there is a finite Markov subsystem $\mathcal{S}' \subset \mathcal{S}$ such that $\gamma_u(\mathcal{S}) = \gamma_u(\mathcal{S}')$.*

Details of this proof are found in [1].

References

[1] I. Dumitriu, P. Tetali, P. Winkler, *On Playing Golf with Two Balls*. Siam J. Discrete Math. 16 (2003), 604-615.

[2] J. Gittins and G. Jones, *A dynamic allocation index for the design of experiments*, Progress in Statistics, Colloq. Math. Soc. János Bolyai 9, J. Gani, K. Sarkadi, and I. Vince, eds., North-Holland, Amsterdam, 1974, pp. 241-266.