On Playing Golf With Two Balls

Ioana Dumitriu [*]　　Prasad Tetali [†]　　Peter Winkler [‡]

January 20, 2003

Abstract

We analyze and solve a game in which a player chooses which of several Markov chains to advance, with the object of minimizing the expected time (or cost) for *one* of the chains to reach a target state. The solution entails computing (in polynomial time) a function γ—a variety of "Gittins index"—on the states of the individual chains, the minimization of which produces an optimal strategy.

It turns out that γ is a useful cousin of the expected hitting time of a Markov chain, but is defined, for example, even for random walks on infinite graphs. We derive the basic properties of γ and consider its values in some natural situations.

1 Introduction

Everyone has encountered situations where there is more than one way to accomplish some task, and where it may be desirable to change strategies from time to time depending on the outcome of various actions. In trying to contact a colleague, for example, one might first try telephoning, and depending on the result, telephone again later or perhaps try sending electronic mail. A dating strategy for someone who is seeking a mate might call for trying a new prospect, or re-trying an old one, if things are going badly with the current one. In these situations, if one knows the best *first* move from any state, one can behave optimally.

Suppose you are invited to play the following game. Tokens begin on vertices 2 and 5 of a path connecting vertices $0, \ldots, 5$ (see Fig. 1). A valuable gift awaits you if either token reaches vertex 3. At any time you may pay \$1 and point to a token; that token will then make a random move (with equal probability to its left or right neighboring vertex, if it has two neighbors, otherwise to its only neighbor). Which token should you move first?

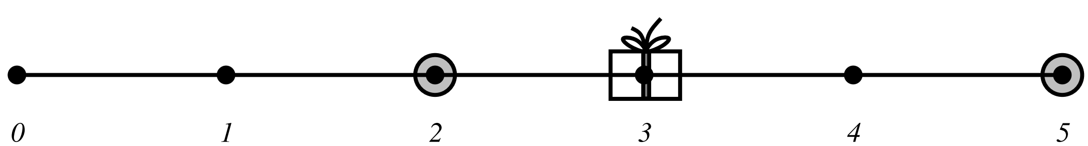

Figure 1: What's the fastest way to the gift?

It is not difficult to see that by moving the token at 2 first, then switching permanently to the other if the game does not end immediately, your expected cost to reach the prize is \$3;

[*]Department of Mathematics, M.I.T., Cambridge, MA 02139
[†]School of Mathematics and College of Computing, Georgia Institute of Technology, Atlanta, GA 30332-0160.
[‡]Bell Labs, Lucent Technologies Inc., 2C-365, Murray Hill, NJ 07974

this is the unique optimal strategy. Contrast this with a similar "loyal" game in which you must choose one token and stick with it. Then, choosing the token at vertex 5 costs \$4 on average, and choosing the other token costs \$5 on average.

Now fix any graph G with distinguished target node t, and write $u \leq v$ if, in the first-described game with tokens at u and v, there is an optimal strategy which calls for moving the token at u first. Is this relation transitive—that is, if $u \leq v$ and $v \leq w$ does that imply $u \leq w$? (Note that two tokens may occupy the same vertex; in fact there is no loss or gain of generality if each token has its own graph and its own target).

In the loyal game, the corresponding statement is trivially true because there is a quantity (the expected length of a random walk from u to t, or the *hitting time*) which measures the desirability of choosing the token at u, regardless of where the other token may be.

When one is permitted to switch tokens the situation becomes more subtle. Nonetheless, it does turn out that there is a measure of first-move desirability which can be applied to a single token, and therefore transitivity does hold. This measure (our function γ) is polynomial-time computable, and it is related both to what Markov decision theorists know as the *Gittins index* of a single-armed bandit, and to expected hitting times in a *different* Markov chain. The development here, however, will be mostly self-contained.

The main theorem will be stated in a more general, but by no means the *most* general, form. The graph is replaced by two (or more) "Markov systems," one for each token; each system consists of a finite-state Markov chain with a starting state and a target state, and a positive real move cost at each state. Further generalizations are considered along the way.

2 Markov Systems

We call the pieces from which we construct our games *Markov systems*. A Markov system $\mathcal{S} = \langle V, P, C, s, t\rangle$ consists of a state space V (which will be assumed finite unless otherwise specified), a transition matrix $P = \{p_{u,v}\}$ indexed by V, a positive real move cost C_v for each state v, a starting state s, and a target state t. We will assume usually that t is accessible (ultimately) from every state in V.

The cost of a "trip" $v(0), \ldots, v(k)$ on $\mathcal{S}$ is the sum $\sum_{i=0}^{k-1} C_{v(i)}$ of the costs of the exited states. The (finite) expected cost of a trip from v to the target t is denoted $\mathrm{E}_v[\mathcal{S}]$, with the subscript sometimes omitted when the trip begins at s. Since we never exit the target state, we may arbitrarily set $C_t = 0$ and $p_{t,t} = 1$.

3 Games and Strategies

The games we consider are all played by a single player against a "bank" and consist of a series of moves chosen and paid for by the player, with random effect. We imagine that the player is forced to play until termination, which occurs mercifully in finite expected time.

The *cost* $\mathbf{E}[\mathcal{G}]$ of a game $\mathcal{G}$ is the minimum expected cost (to the player) of playing $\mathcal{G}$, taken over all possible strategies. A strategy which achieves expected cost $\mathbf{E}[\mathcal{G}]$ is said to be *optimal*.

Let $\mathcal{S}_1, \ldots, \mathcal{S}_k$ be k Markov systems, each of which has a token on its starting state, and an associated cost function C_i. A *simple multi-token Markov game* $\mathcal{S}_1 \circ \mathcal{S}_2 \circ \cdots \circ \mathcal{S}_k$ consists of a succession of steps in which we choose one of the k tokens, which takes a random step in its system (i.e., according to its P_i). After choosing a token i (on state u, say), we pay the cost $C_i(u)$ associated with the state u of the system $\mathcal{S}_i$ whose token we have chosen. As soon as one of the tokens reaches its target state for the first time, we stop.

We define the *terminator* $\mathcal{T}_g$ as the Markov system $\langle\{s,t\},P,g,s,t\rangle$ where $p_{s,t}=1$. The terminator always hits its target in exactly one step, at cost g. The *Token vs. Terminator* game, in which we play a simple two-token game of systems $\mathcal{S}$ (for some $\mathcal{S}$) and $\mathcal{T}_g$ (for some g), will play a critical role, in the analysis of general Markov games.

It will also be useful to define the *join* $\mathcal{G}=\mathcal{G}_1\Box\mathcal{G}_2\Box\cdots\Box\mathcal{G}_n$ of games $\mathcal{G}_1,\ldots,\mathcal{G}_n$, as follows: at each step the player chooses one of the n games, then pays for and makes a move in that game. $\mathcal{G}$ terminates when any of its component games is finished. We will employ the join in order to analyze the Markov game $\mathcal{S}_1\circ\mathcal{S}_2\circ\cdots\circ\mathcal{S}_k$.

Throughout the paper, we will be using (sometimes without making explicit reference to) the following two classical theorems from the general theory of Markov game strategies; the reader is referred to [15] for more detail.

The first theorem enables us to look for optimal strategies in a finite set.

Theorem 3.1. *Every Markov game (in our sense) has a pure optimal strategy.*

From a given *state* u of a Markov game, an *action* α produces an immediate expected cost $C_u(\alpha)$, and a probability distribution $\{p_{u,\cdot}\}$ of new states. (Note that in the present context, a *state* of a Markov game consisting of k Markov systems would be a specific configuration of the k tokens, and an *action* would correspond to the choice of a particular token to move.) thus a strategy σ which takes action α at the state u satisfies

$$E_u[\sigma]=C_u(\alpha)+\sum_v p_{u,v}(\alpha)E_v[\sigma]\ .$$

If among all possible actions at state u, α minimizes the right-hand side of this expression, σ is said to be *consistent* at u.

Theorem 3.2. *A strategy σ is optimal iff it is consistent at every state.*

Proof. Let τ be an optimal strategy, and U the set of states v on which $E_v[\tau]-E_v[\sigma]$ attains its minimal value x. Since $t\notin U$, we can find a state $u\in U$ from which some state not in U can be reached in one step by τ. But then, if α is the action taken by τ at u,

$$E_u[\sigma]=E_u[\tau]+x=C_u(\alpha)+\sum_v p_{u,v}(\alpha)(E_v[\tau]+x)<C_u(\alpha)+\sum_v p_{u,v}(\alpha)E_v[\sigma],$$

contradicting the fact that σ is consistent at u. □

4 The Grade

We say that an optimal player is *indifferent* among some set of moves if for each of those moves there is an optimal strategy which employs it. Going back to the *Token vs. Terminator* game from the preceding section, we define the *grade* $\gamma(\mathcal{S})$ of a system $\mathcal{S}=\langle V,P,C,s,t\rangle$ to be the unique value of g at which an optimal player is indifferent between the two possible first moves in the game $\mathcal{G}_g=\mathcal{S}\circ\mathcal{T}_g$. Thus, $\gamma(\mathcal{S})$ is the least value of g such that if, at any time, we can pay g to quit the system $\mathcal{S}$, we are still willing to try one move in $\mathcal{S}$. (To be consistent with our notation below, we should be denoting $\gamma(\mathcal{S})$ by $\gamma_s(\mathcal{S})$, indicating the start state of the game.)

To see that $\gamma=\gamma(\mathcal{S})$ is a well-defined quantity, we will make use of Theorem 3.1. Any pure strategy σ is defined by the set $Q(\subset V)$ of states in which it chooses to move in $\mathcal{T}_g$. Suppose

$\mathcal{S}$ is run until either t or a state in Q is reached; let the first event be represented by R, and let X be the final cost of the run in $\mathcal{S}$. Put $p = \Pr[R]$, $A = \mathrm{E}[X|R]$, and $B = \mathrm{E}[X|\neg R]$. Then

$$\mathrm{E}[\sigma] = pA + (1-p)(B+g) \ ,$$

i.e. $E[\sigma]$ is linear in g for fixed strategy σ. Optimizing over σ, it follows that $\mathbf{E}[\mathcal{G}_g]$ is the maximum of a set of linear functions, therefore continuous in g. For g less than the cost of the first move, $\mathbf{E}[\mathcal{G}_g] = g$ (because we choose to move in $\mathcal{T}_g$). On the other hand if g exceeds the expected cost $\mathrm{E}_v[\mathcal{S}]$ from any state v, then we will always choose to move in $\mathcal{S}$, hence $\mathbf{E}[\mathcal{G}_g] = \mathrm{E}_s[\mathcal{S}]$. Fig 2 illustrates a typical shape for the graph of $\mathbf{E}[\mathcal{G}_g]$.

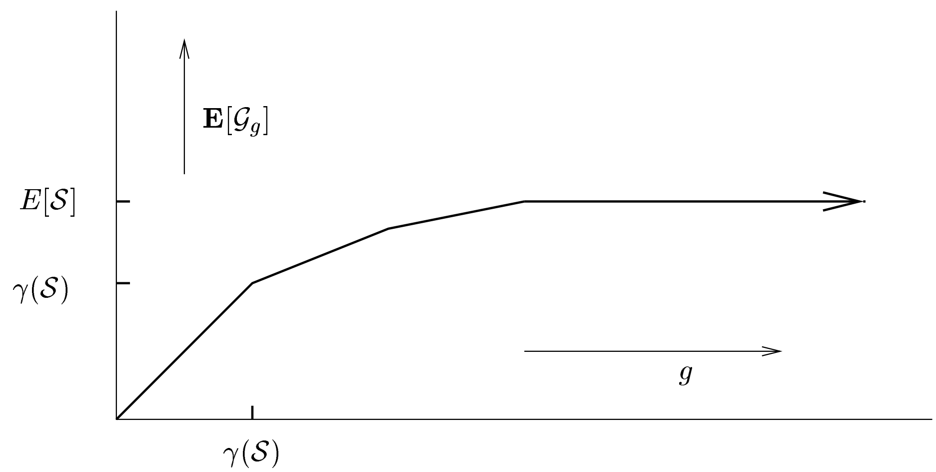

Figure 2: The expected cost of $\mathcal{G}_g = \mathcal{S} \circ \mathcal{T}_g$

The grade of $\mathcal{S}$ is marked on the figure as the highest value of g at which the strategy "play in $\mathcal{T}_g$" is optimal, i.e. the coordinate of the top end of the line segment of slope 1.

We use $\gamma_u(\mathcal{S})$ or just γ_u (when $\mathcal{S}$ is fixed except for its starting state) to denote the grade of $\mathcal{S}_u = \langle V, P, C, u, t\rangle$. Hence we can formulate the following theorem:

Theorem 4.1. *A strategy for $\mathcal{S} \circ \mathcal{T}_g$ is optimal iff it chooses $\mathcal{S}$ whenever the current state u of $\mathcal{S}$ satisfies $\gamma_u < g$ and it chooses $\mathcal{T}_g$ whenever $\gamma_u > g$.*

Remark 4.2. Note that in system $\mathcal{S} = \langle V, P, C, s, t\rangle$ there is positive probability of moving from s to a state of strictly lower grade. Otherwise, in the "token vs. terminator" game $\mathcal{S} \circ \mathcal{T}(\gamma(\mathcal{S}))$ the strategy of paying for the first move in $\mathcal{S}$ and then terminating would be optimal, yet more costly than terminating immediately.

5 An Optimal Strategy for the Simple Multitoken Game

The surprising and fundamental discovery of John Gittins, first proved by Gittins and Jones [6], was that in many Markov games, options could be "indexed" separately and then numerically compared to determine an optimal strategy. This is indeed the case for our games, the index being the "grade" defined above.

Theorem 5.1. *A strategy for the game $\mathcal{G} = \langle \mathcal{S}_1 \circ \cdots \circ \mathcal{S}_n\rangle$ is optimal if and only if it always plays in a system whose current grade is minimal.*

Proof. We will employ a modified version of the very elegant proof given by Richard Weber [14] for Gittins' theorem. Our "grade" differs from the Gittins index in several minor respects, among them that our games terminate and our costs are not subject to discounting (about which more later). These differences are not sufficient to regard the grade as other than a special case, or variation, of the Gittins index.

The proof will proceed using a sequence of easy lemmas. We begin by considering a "reward game" $\mathcal{S}_i(g)$ based on the system $\mathcal{S}_i$, in which we play and pay as in $\mathcal{S}$ but may quit at any time; as incentive to play, however, there is a reward of g at the target which we may claim when and if the target is reached.

Lemma 5.2. *$\mathcal{S}_i(\gamma(\mathcal{S}_i))$ is a fair game (that is, the expectation $\mathbf{E}[\mathcal{S}_i(\gamma(\mathcal{S}_i))] = 0$) and a strategy for $\mathcal{S}_i(\gamma(\mathcal{S}_i))$ is optimal if and only if the player quits whenever his current state u satisfies $\gamma_u > g$, and plays on when $\gamma_u < g$.*

Proof. The reward game is no different from a terminator game $\mathcal{S} \circ \mathcal{T}_{\gamma(\mathcal{S}_i)}$ in which the player is provided with an initial stake of $\gamma(\mathcal{S}_i)$, hence the characterization of optimality follows from Theorem 4.1. Since quitting immediately is among the optimal strategies, $\mathbf{E}[\mathcal{S}_i(\gamma(\mathcal{S}_i))] = 0$. $\square$

Suppose the game $\mathcal{S}_i(\gamma_u(\mathcal{S}_i))$ is amended in the following teasing manner: whenever the player reaches a state u with $\gamma_u > g$, the reward at the target is boosted up to γ_u—just enough to tempt the player to continue. (Note that the reward is never lowered.) It might seem that this game, which we will denote simply by $\mathcal{S}'_i$, is better than fair, but:

Lemma 5.3. *$\mathcal{S}'_i$ is fair, and a strategy for $\mathcal{S}'_i$ is optimal if and only if the player never quits when the current grade is below the current reward value.*

Proof. To see that $\mathbf{E}[\mathcal{S}'_i] = 0$, note that a session with $\mathcal{S}'_i$ can be broken up into a series of smaller games, each ending when either at a state u whose grade exceeds the current reward value, or at the final target. Since each of these games is fair, so is $\mathcal{S}'_i$. Note that the antipodal strategies of quitting immediately, and of playing until the target is hit, are in particular both optimal. $\square$

Now we consider the join $\mathcal{G}' := \mathcal{S}'_1 \square \cdots \square \mathcal{S}'_n$, in which we play the teaser game of our choice, paying as we go, until we quit or hit one of the targets (in which case we claim the current reward at that target).

Lemma 5.4. *$\mathcal{G}'$ is a fair game.*

Proof. Any combination (simultaneous, sequential, or interleaved) of the independent fair games $\mathcal{S}'_1, \dots, \mathcal{S}'_n$ is still fair. The join $\mathcal{G}'$ can be no better than such a combination, since it differs only in having additional restrictions on the player; hence it is *at best* fair. However, $\mathcal{G}'$ cannot be worse than fair, since e.g. the player can simply quit at the start or play one game to its finish and ignore the others. $\square$

Among the strategies for $\mathcal{G}'$ is one we call the "Gittins strategy" Γ: always play from a system which is currently of minimal grade. This is the strategy we claim is optimal for the original game $\mathcal{G}$, but first we observe two properties of Γ relative to the game $\mathcal{G}'$.

Lemma 5.5. *The Gittins strategy Γ is optimal for $\mathcal{G}'$.*

Proof. If a move by Γ results in the grade of a component game $\mathcal{S}'_i$ dropping below its reward value, then since its grade has just gone down it is now the *unique* lowest-grade component and therefore Γ will again move that token. Hence no component system will ever be stranded in a state u with γ_u less than the reward on target t_i, thus all the components $\mathcal{S}'_i$ are played optimally. $\square$

Lemma 5.6. *Of all strategies for $\mathcal{G}'$ which play until a target is hit, Γ reaps the smallest expected reward at the end. In other words, if the move-costs are waived, then Γ actually is the* worst *(in terms of reward collected) possible non-quitting strategy for $\mathcal{G}'$. Furthermore, among non-quitting strategies which are optimal for the unaltered game $\mathcal{G}'$, Γ is the only one with this property.*

Proof. Imagine that the course of each individual system $\mathcal{S}_i$ is fixed. Then each teaser game $\mathcal{S}'_i$ terminates, if played all the way to its target, with a certain reward g_i (equal to the largest γ_u over all states u hit en route). Every non-quitting strategy will claim one of the rewards g_i at the end, but the Gittins strategy gets the *smallest* one; the reason is that if it collected a non-minimal reward (say g_j) when teaser game $\mathcal{S}'_i$, $i \neq j$, was headed for a final reward of $g_i < g_j$, then at the time of termination of $\mathcal{G}'$ the reward for $\mathcal{S}'_i$ was g_i or less, hence $\gamma_u(\mathcal{S}_i) \leq g_i$ where u is its last state. But this is impossible because the final run of plays of $\mathcal{S}_j$ began at a state v where $\gamma_v(\mathcal{S}_j) = g_j$ or more, and Γ should have preferred to play in $\mathcal{S}_i$ at that time. From the proof it is clear that Γ is unique in the sense of the last assertion if the statement of the lemma. $\square$

We are finally ready to show that Γ is optimal for the original game $\mathcal{G}$. For any non-quitting strategy Δ for $\mathcal{G}'$, let $C(\Delta)$ be its expected cost and $R(\Delta)$ its expected reward; thus $\mathrm{E}[\Delta] = R(\Delta) - C(\Delta) \leq 0$ since $\mathcal{G}'$ is fair. But then since $\mathrm{E}[\Gamma] = 0$,

$$C(\Gamma) = R(\Gamma) \leq R(\Delta) \leq C(\Delta)$$

so Γ incurs the least cost among all non-quitting strategies for $\mathcal{G}'$, and this says exactly that it is optimal for $\mathcal{G}$.

If Δ is also optimal for $\mathcal{G}$ then the above inequalities are both tight, hence Lemmas 5.5 and 5.6 both hold for Δ. If Δ is not a Gittins strategy then we may assume that Δ makes a non-Gittins move already at the start of the game, playing $\mathcal{S}_2$ even though $\mathcal{S}_1$ has smaller grade. This will not necessarily cause it to miss the smallest reward in $\mathcal{G}'$, because there may be 0 probability of that system hitting its target immediately and Δ can return to $\mathcal{S}_1$ before it's too late. However, it follows from Remark 4.2 above that there is *always* a positive probability that any system will reach its target along a path whose grade is strictly declining. If this is fated to happen to both $\mathcal{S}_1$ and $\mathcal{S}_2$ then Δ will either end up accepting the larger reward of $\mathcal{S}_2$ (thus failing to have minimal reward) or leave one of the systems in a "grade below reward" state (thus failing to be optimal for $\mathcal{G}'$).

We conclude that Δ is optimal for $\mathcal{G}$ if and only if it is a Gittins strategy, and the proof of Theorem 5.1 is complete. $\square$

6 The Grade and the Gittins Index

Both the history and the range of applicability of the Gittins index are rather complex subjects; the reader is referred to Gittins' modestly-written book [5] for some appreciation of the former. It appears that the mathematical and statistical communities took some time to appreciate that the notorious "multi-armed bandit" problem had been solved, then additional time to find new, cleaner proofs and to uncover some very nice disguised consequences. The experience of this paper's authors suggests that the Gittins index is still not widely known in the mathematical community, especially among researchers in combinatorics and in the theory of computing. We hope to make a start at rectifying the situation with this work.

Framed in our terms, the circumstances to which the Gittins index was originally applied comprise a collection of Markov systems $\mathcal{S}_1, \dots, \mathcal{S}_n$ such as those we have considered but without target states, and with rewards instead of costs. When a system is chosen (say at time t) a (possibly random) non-negative and uniformly bounded reward R_t, dependent on the state of that system, is collected. The object is to maximize $\sum_{t=0}^{\infty} \beta^t R_t$ where β is a "discount" strictly between 0 and 1.

"Gittins' theorem" asserts the existence of an index depending on system and state whose maximization at each stage produces an optimal strategy. Readers are referred to [16] and [12] as well as [6], [14] and Gittins' book [5] for various proofs.

The discount β is ubiquitous in Markov decision theory and in economics, financial and actuarial research as well; and it is necessary in the multi-armed bandit formulation to make the objective function finite. Discounts are less natural and familiar to pure mathematicians and are obviated in our presentation, where the presence of terminating targets keeps things finite. The elimination of discounts, particularly in the context of job scheduling problems, is discussed in Section 6.2 of [5]; one approach, which can be used to deduce Theorem 5.1 from Gittins' theorem, is to let targets represent cycling states of zero reward and allow the discount factor to approach 1.

One benefit of our formulation is its natural application to the problem of minimizing the time needed to reach a goal, for example for some token on a graph to reach a target node via random walk. As a result we can represent our "grade" γ as a hitting time, or more generally a hitting cost.

Let $\mathcal{S} = \langle V, P, C, s, t \rangle$ be a Markov system and let $U \subset V \setminus \{s, t\}$. Define a new system $\mathcal{S} \restriction U = \langle U \cup \{s, t\}, P', C, s, t \rangle$ by putting

$$p'_{u,s} = p_{u,s} + \sum_{v \in V \setminus \{U \cup \{s,t\}\}} p_{u,v}$$

and $p'_{u,w} = p_{u,w}$ for $w \in U$ and $u \in U$. In effect, $\mathcal{S} \restriction U$ is the restriction of $\mathcal{S}$ to U where the state-marking token is sent back to s whenever it tries to leave U.

Theorem 6.1. *With $\mathcal{S}$ and U as above, $\gamma_s(\mathcal{S}) \le \mathrm{E}_s[\mathcal{S} \restriction U]$, with equality if U contains all states of grade lower than γ_s and no states of grade higher than γ_s.*

Proof. Let σ be the strategy for playing the "reward game" $\mathcal{S}(\gamma_s(\mathcal{S}))$ which entails playing until the target is hit or some state $v \in V \setminus U$ is reached, in which case the game is terminated. Since $\mathcal{S}(\gamma_s(\mathcal{S}))$ is a fair game, σ has non-positive expectation. Suppose we are permitted to restart a new $\mathcal{S}(\gamma_s(\mathcal{S}))$, and continue with strategy σ, whenever there is a voluntary termination. The resulting sequence of games still has non-positive expectation, but is equivalent to playing the reward game $\mathcal{S} \restriction U(\gamma_s(\mathcal{S}))$ until the target is hit. Since this will always result in collecting $\gamma_s(\mathcal{S})$ at the end, the expected total move-cost must be at least $\gamma_s(\mathcal{S})$.

On the other hand, we know from Theorem 4.1 that σ is optimal (thus has zero expected reward) when U fulfills the additional conditions; in that case we get that the expected total move-cost is precisely $\gamma_s(\mathcal{S})$. ∎

Note that the $U = \emptyset$ case yields the rather obvious fact that $\gamma_x \leq \mathrm{E}_x[\mathcal{S}]$ for all x.

It might be argued that Theorem 6.1 is circular since it reduces computing the grade to computing a hitting cost, but only if we know which states have grade less than γ_s, and which more. However, in the next section we use the theorem recursively to compute grades one by one.

7 Computing the Grade

Like (most variations of) the Gittins index, our "grade" can be determined in time bounded by a polynomial in the length of description of a system $\mathcal{S}$. We will now present, and analyze, an algorithm which calculates the grade γ_u of all the states u of $\mathcal{S}$, one state at a time.

Let U be the set of states in V whose grades have already been calculated. We add one more state to U, namely, the state of smallest grade in $V \backslash U$. Let $N(U)$ denote the set of states x in $V \backslash U$ that are reachable directly from a state in U (i.e., $N(U) := \{v \in V | p_{u,v} > 0$ for some $u \in U\}$).

As before, $\mathrm{E}_x[\mathcal{S}]$—the "hitting cost"—denotes the expected cost of a trip to t from x.

The algorithm is given in pseudo-code below.

1. $U = \{t\},\ \gamma_t = 0;$
2. While $V \backslash U \neq \phi$
   (a) *CheckedStates* $= \phi$;
   (b) While *CheckedStates* $\neq N(U)$
      i. Choose $v \in N(U) \backslash$ *CheckedStates*;
      ii. Let $P' = \{p'_{u,v}\}$ be the transition matrix obtained from $P = \{p_{u,v}\}$ in the following way:
         - P' disregards all states not in $U \cup \{v\}$;
         - $p'_{u,v} = p_{u,v} + \sum_{w \in V \backslash \{U \cup \{v\}\}} p_{u,w},\ \forall u \in U \cup \{v\};$
         - $p'_{u,u'} = p_{u,u'},\ \forall u \in U \cup \{v\},$ and $u' \in U$.
      iii. Compute $h_v = \mathrm{E}_v[\mathcal{S}']$ where $\mathcal{S}' = \langle U, P', C, v, t \rangle$;
      iv. *CheckedStates* = *CheckedStates* $\cup\{v\}$;
   (c) Find x such that $h_x = \min\{h_v : v \in$ *CheckedStates*$\}$;
   (d) $U = U \cup \{x\},\ \gamma(x) = h_x$.

It is evident from Theorem 6.1 that if the selected state x always has minimum grade among the states in $V \backslash U$, then the algorithm correctly computes the grades of all states in V.

We first note that a minimum-grade $x \notin U$ is indeed to be found among the neighbors of U, because Remark 4.2 implies that there is a path of decreasing grade from x to t.

It remains only to establish that if $v \in V \backslash U$ is *not* of minimum grade, then $h_v = \mathrm{E}_v[\mathcal{S}']$ is at least as large as γ_v. But this is exactly the content of Theorem 6.1 of the preceding section.

Let us now analyze the running time for the algorithm.

Let n be the initial number of states. At step i, $N(U)$ has $\mathrm{O}(n-i)$ states. For any state in $N(U)$, the greatest workload is done to compute $E_v(P')$. It involves solving an $(i+1) \times (i+1)$ system of equations; this can be done by an LU factorization followed by a backward substitution, and it represents $\mathrm{O}(i^3)$ work. Therefore, we can compute all the grades in $\mathrm{O}(\sum_{i=1}^{n}(n-i)i^3)$ $= \mathrm{O}(n^5)$ time.

8 States of Maximum Grade

If the starting state of system $\mathcal{S}$ has maximum grade, then "never quitting" is an optimal strategy for the Token vs. Terminator game $\mathcal{S} \circ \mathcal{T}_\gamma$. Hence

Lemma 8.1. *Let z be a state of maximum grade in a system $\mathcal{S}$. Then $\gamma_z = \mathrm{E}_z[\mathcal{S}]$.*

The converse of Lemma 8.1 fails for the uninteresting reason that states of higher grade than z may exist but not be accessible from z. More interesting is the question of maximum grade versus maximum hitting cost (that is, maximum expected cost of hitting t).

Theorem 8.2. *In any system $\mathcal{S}$ the states of maximum grade and the states of maximum hitting cost are the same.*

Proof. Suppose that x maximizes $\mathrm{E}_x[\mathcal{S}]$, that is, x incurs the greatest expected cost $h_x = \mathrm{E}_x[\mathcal{S}]$ of hitting t assuming best strategy. Then we claim that $\gamma_x = h_x$. To see this, we let U be the set of states u in V such that $\gamma_u > \gamma_x$ and compute h_x by considering the effect of the event "A" that a walk from x hits U before it reaches t. Then

$$\begin{aligned} h_x &= \Pr[\neg A]\mathrm{E}_x[\mathcal{S}|\neg A] + \Pr[A]\left(\mathrm{E}_x[U] + \mathrm{E}_U[\mathcal{S}]\right) \\ &\leq \Pr[\neg A]\mathrm{E}_x[\mathcal{S}|\neg A] + \Pr[A]\left(\mathrm{E}_x[U] + h_x\right) \end{aligned}$$

where $\mathrm{E}_x[U]$ is the expected cost of hitting U from x and $\mathrm{E}_U[\mathcal{S}]$ is the expected cost of hitting t from the random point in U which is hit first. Solving, we get

$$h_x(1 - \Pr[A]) \leq \Pr[\neg A]\mathrm{E}_x[\mathcal{S}|\neg A] + \Pr[A]\mathrm{E}_x[U]\ .$$

However, if we compute $\gamma_x = \mathrm{E}_x[\mathcal{S} \restriction U]$ in the same fashion, we get

$$\mathrm{E}_x[\mathcal{S} \restriction U](1 - \Pr[A]) = \Pr[\neg A]\mathrm{E}_x[\mathcal{S}|\neg A] + \Pr[A]\mathrm{E}_x[U]$$

so that $h_x \leq \gamma_x$; thus they are equal. In particular, $\gamma_y \leq h_y \leq h_x = \gamma_x$ for all y so x also has maximal grade.

Suppose on the other hand that z has maximal grade, but not maximal hitting cost; let x have maximal hitting cost. But then we have seen that $\gamma_x = h_x > h_z \geq \gamma_z$, a contradiction. The theorem follows. □

Remark 8.3. Theorems 6.1 and 8.2 provide an algorithm for computing grades from highest to lowest, as opposed to the one we presented earlier. The idea is to find the state x_1 of largest hitting cost (hence highest grade), then the state x_2 which maximizes $\mathrm{E}_{x_2}[\mathcal{S} \restriction (V \setminus \{x_1\})]$ etc. Although we are not able to take advantage here of the neighborhood structure, the running time for this algorithm is of the same order as before, relative to the number of states.

9 Grades and Graphs

The hitting time (from, say, x to y) for a simple random walk on a graph G has many beautiful properties, including ties to electrical networks; our analogue, the "grade", has the additional advantage of being finite even when G is infinite. Below we illustrate some calculations and theorems concerning the grades of vertices of some symmetric graphs.

We have assumed up until now that our Markov chains have finite state spaces, and indeed it would appear that there are problems with the expected outcome of our basic game when the expected number of steps to hit a target is infinite; or, even worse, when there is positive probability that the target will never be hit. However, the simple multi-token game makes sense as long as at least one of the systems it deals with has a finite hitting time to the target, and of course the "terminator" system has this property. It is not difficult to prove that:

Theorem 9.1. *Let $\mathcal{M}$ be an infinite, locally finite Markov chain, with designated target state t. Then*

1. *every state u of $\mathcal{M}$ has a grade $\gamma_u = \gamma_u(\mathcal{M}) < \infty$;*
2. *for all real k, the set $S_k = \{v \in \mathcal{M} : \gamma_v < k\}$ is finite;*
3. *for all $u \in \mathcal{M}$, there exists a finite chain $\mathcal{M}'$, obtained via suppressing all but a finite number of states in $\mathcal{M}$, for which $\gamma_u(\mathcal{M}) = \gamma_u(\mathcal{M}')$.*

We will sketch the proof of Theorem 9.1; it is left to the reader to fill in the details.

Proof. 1. Since the Markov chain is locally finite, it follows that from any state u there is a (finite) shortest path to t. Let u be an arbitrary state, let $k < \infty$ be the length of a shortest path from u to t, and let p be the probability of this path (since the chain is locally finite, it follows that $p > 0$). Let $g_* = k/p$, and consider the Token vs. Terminator $\mathcal{T}_{g_*}$ game, with the following strategy: starting from u, move k times "blindly" on $\mathcal{M}$, paying before each move; if after k steps the token is not on the target, pay the Terminator and end the game in a step.

   It is immediate to verify that this strategy, though perhaps suboptimal, breaks even: the expected profit/loss from it is 0. But if γ_u were infinite, then for any finite g (in particular for g_*), any strategy for the Token vs. Terminator T_g game that does not choose the Terminator T_g immediately would guarantee a positive loss! Hence γ_u must be finite. More over, it also follows that $g_* \geq \gamma_u$.

2. A state v whose distance from t is at least k will also (necessarily) have a grade of at least k; this is equivalent to saying that for any real k, $S_k \subseteq D_k = \{v \in \mathcal{M} : dist(v,t) < k\}$. Due to the local finiteness of the chain, for any real k, the set D_k is finite; hence for any real k, S_k is finite.

3. This follows directly from 1 and 2: given a state u, let $k = \gamma_u$, and suppress all states of $\mathcal{M}$ but for those in S_k. In the newly obtained finite chain $\mathcal{M}'$, $\gamma_u(\mathcal{M}) = \gamma_u(\mathcal{M}')$.

□

In the following subsections, we consider the grade function for the simple random walk on each of the following graphs: the hypercube, the Cayley tree, the plane square grid and the cubic grid in three-space. The last three are immediately relevant to the above, as infinite, locally finite chains; the first is finite, but interesting in itself.

9.1 The Hypercube

We begin with a finite graph, the n-dimensional hypercube Q^n, whose vertices are binary sequences $u = (u_1, \ldots, u_n)$ with $u \sim v$ when they differ in just one coordinate. The "k-th level" of Q^n consists of the vertices with exactly k 1's. If the target vertex is fixed at the origin, then the grade γ_k of a point in level k is the hitting time from level k to level 0 in the truncated hypercube Q_k^n, defined as follows: all vertices at level greater than k are deleted, and each vertex at level k is provided with $n-k$ loops so that its total degree is n.

Let T_j be the time it takes to get from level j to level $j-1$ in Q_k^n. Clearly $\gamma_k = \sum_{j=1}^{k} T_j$; we derive the following recursion for T_j:

$$\begin{aligned} T_k &= \frac{n}{k}, \quad \text{and} \\ T_j &= 1 + \frac{n-j}{n}\Big(T_j + T_{j+1}\Big). \end{aligned}$$

It is straightforward to verify that

$$T_j = \frac{1}{\binom{n-1}{n-j}} \sum_{i=n-j}^{n} \binom{n}{i};$$

this yields

$$\gamma_k = \sum_{i=1}^{k} \frac{1}{\binom{n-1}{n-j}} \sum_{i=n-j}^{n} \binom{n}{i}.$$

9.2 The Cayley Tree

The d-regular Cayley tree is the unique connected, cycle-free infinite graph T^d each of whose vertices has degree d. Again, this is a symmetric graph so we may assume the target vertex is an arbitrary "root" t.

The case $d = 2$ is the doubly-infinite path, in which the grade of a vertex v at distance k from t is easily seen to be $k(k+1)$.

In general the grade γ_k of a vertex v at distance k from the root is the hitting time from v to t in the graph T_k^d consisting of the first k levels of the tree (the root being at level 0), in which every vertex on the last level has $d-1$ loops (instead of $d-1$ children). This leads to a recurrence to which the solution, for $d > 2$, is:

$$\gamma_k = \frac{d\Big((d-1)^{k+1} - 1 - (k+1)(d-2)\Big)}{(d-2)^2}. \tag{1}$$

Interestingly, there is another way to compute the grade on T^d which works on any finite tree and shows that on trees, grades and hitting times are always integers. Let T be any tree, possibly with loops. Fix a target vertex t, and let v any other vertex. Order the edges (including loops) incident to each $u \neq t$ arbitrarily subject to the edge on the path from u to t being last. Now walk from v by choosing each exiting edge in round-robin fashion, in accordance with the edge-order at the current vertex, until t is reached. For example, if the edges incident to some degree-3 vertex u are ordered e_1, e_2, e_3, then the first time u is reached it is exited via e_1, the second time by e_2, the fourth time by e_1 again, etc. We call such a walk a "whirling tour"; an example is provided in figure 3.

Theorem 9.2. *In any finite tree (possibly with some loops) the length of any whirling tour from v to t is exactly the expected hitting time from v to t.*

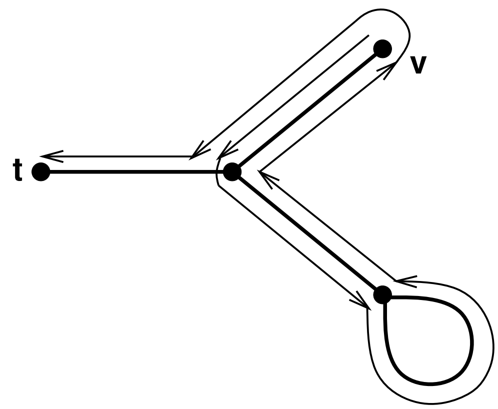

Figure 3: A whirling tour

We leave the proof to the amusement of the reader.

We will denote by g_k the length of such a walk from the k-th level to the root, for every $k \in \mathbb{N}, k \geq 1$.

In order to walk from level k to level 0 (the root), we have to first execute a walk from level k to level 1, and then walk from there to the root. The rest of the walk will be a depth-first search of a (d–1)-ary tree with loops for leaves, plus the final edge. The length of the depth-first search is easily computed: we have

$$\sum_{i=1}^{k-2}(d-1)^i = \frac{(d-1)^{k-1}-1}{d-2} - 1$$

edges, and $(d-1)^{k-1}$ loops; each edge is walked twice (once forward, once backward) and each loop is walked once, for a total length of

$$2\Big(\frac{(d-1)^k - 1}{d-2} - 1\Big) + (d-1)^k = \frac{d(d-1)^k - 2d}{d-2}.$$

This sets up the recurrence

$$g_k = \frac{d(d-1)^k - 2d + 2}{d-2} + g_{k-1} + 1,$$

where $g_0 = 0,\ g_1 = d$. Thus,

$$g_k = \sum_{j=1}^{n} \frac{d\Big((d-1)^j - 1\Big)}{d-2} = \frac{d\Big((d-1)^{k+1} - 1 - (k+1)(d-2)\Big)}{(d-2)^2},$$

in accordance with (1), as expected.

9.3 Grids

The d-dimensional grid $\mathbb{Z}^d$ is the graph whose vertices are d-tuples of integers, with $u \sim v$ if u and v are at Euclidean distance 1. Since simple random walks on $\mathbb{Z}^d$ behave approximately symmetrically with respect to rotation, one would expect that the Gittins index of a node of

$\mathbb{Z}^d$, with the origin as target, depends largely on its distance from the origin. This and more has recently been verified by Svante Janson and Yuval Peres [8]; we quote their results below. To prove these, Janson and Peres employ a general lemma bounding the grade of each state of a Markov chain on both sides. The bounds are provided by integrals which depend on some harmonic function defined on the states.

Theorem 9.3. *For simple random walk on $\mathbb{Z}^2$,*

$$\gamma(x,0) = 2|x|^2 \ln|x| + (2\gamma + 3\ln 2 - 1)|x|^2 + O(|x|\ln|x|), \qquad |x| \geq 2,$$

where γ on the right hand side is Euler's constant, $\lim_{n\to\infty}(-\log_e n + \sum_{i=1}^n 1/i)$.

Theorem 9.4. *For simple random walk on $\mathbb{Z}^d$, $d \geq 3$,*

$$\gamma(x,0) = \frac{\omega_d}{p_d}|x|^d + O(|x|^{d-1}),$$

where $\omega_d = \pi^{d/2}/\Gamma(d/2+1)$ is the volume of the unit ball in $\mathbb{R}^d$ and p_d is the escape probability of the simple random walk, i.e. the probability that the random walk never returns to its starting point.

From these theorems it follows that for each dimension d there is a constant $C = C(d)$, independent of the starting position x, such that the optimal strategy is to restart from every position y with $|y| > |x| + C$, but never when $|y| < |x| + C$.

References

[1] R. Bellman, *Dynamic Programming*, Princeton University Press, Princeton NJ 1957.

[2] D. Coppersmith, P. Tetali and P. Winkler, Collisions among random walks on a graph, *SIAM J. Disc. Math.* **6** #3 (Aug. 1993), 363–374.

[3] C. Derman, *Finite State Markovian Decision Processes*, Academic Press, New York 1970.

[4] A.V. Gheorghe, *Decision Processes in Dynamic Probabilistic Systems*, Kluwer Academic Publishers, Dordrecht 1990.

[5] J.C. Gittins, *Multi-Armed Bandit Allocation Indices*, Wiley, New York (1989).

[6] J.C. Gittins and D.M. Jones, A dynamic allocation index for the design of experiments, in *Progress in Statistics* (j. Gani, K. Sarkadi and I. Vince, eds.), *Colloq. Math. Soc. János Bolyai* **9**, North-Holland, Amsterdam (1974), 241–266.

[7] R.A. Howard, *Dynamic Probabilistic Systems* (Vols. I & II), John Wiley & Sons, New York 1971.

[8] S. Janson and Y. Peres, Hitting times for random walks with restarts, preprint (2001).

[9] P. Tetali, Random walks and effective resistance of networks, *J. Theoretical Prob.* No. 1 (1991), 101–109.

[10] P. Tetali, An extension of Foster's network theorem, in *Combinatorics, Geometry and Probability*, Cambridge Univ. Press, Cambridge, 1997, 541–547.

[11] P. Tetali and P. Winkler, On a random walk problem arising in self-stabilizing token management, *Proc. 10th ACM Symp. on the Principles of Distributed Computing* (Montreal, 1991), ACM Press, 273–280.

[12] J.N. Tsitsiklis, A short proof of the Gittins index theorem, *Ann. Appl. Prob.* **4** #1 (1994), 194–199.

[13] P. Tetali and P. Winkler, Simultaneous reversible Markov chains, in *Combinatorics, Paul Erdős is Eighty* Vol. 1, D. Miklós, V.T. Sós and T. Szőnyi, eds, János Bolyai Mathematics Institute, Budapest (1993), 433–452.

[14] R. Weber, On the Gittins index for multiarmed bandits, *Ann. Appl. Prob.* **2** #4 (1992), 1024–1033.

[15] D.J. White, *Markov Decision Processes*, John Wiley & Sons, New York 1993.

[16] P. Whittle, Multi-armed bandits and the Gittins index, *J. Royal Stat. Soc. London* (Series B) **42** #2 (1980), 143–149.