Modeling the Tennis Match as a Markov Chain

Vasco Villas-Boas

September 5, 2022

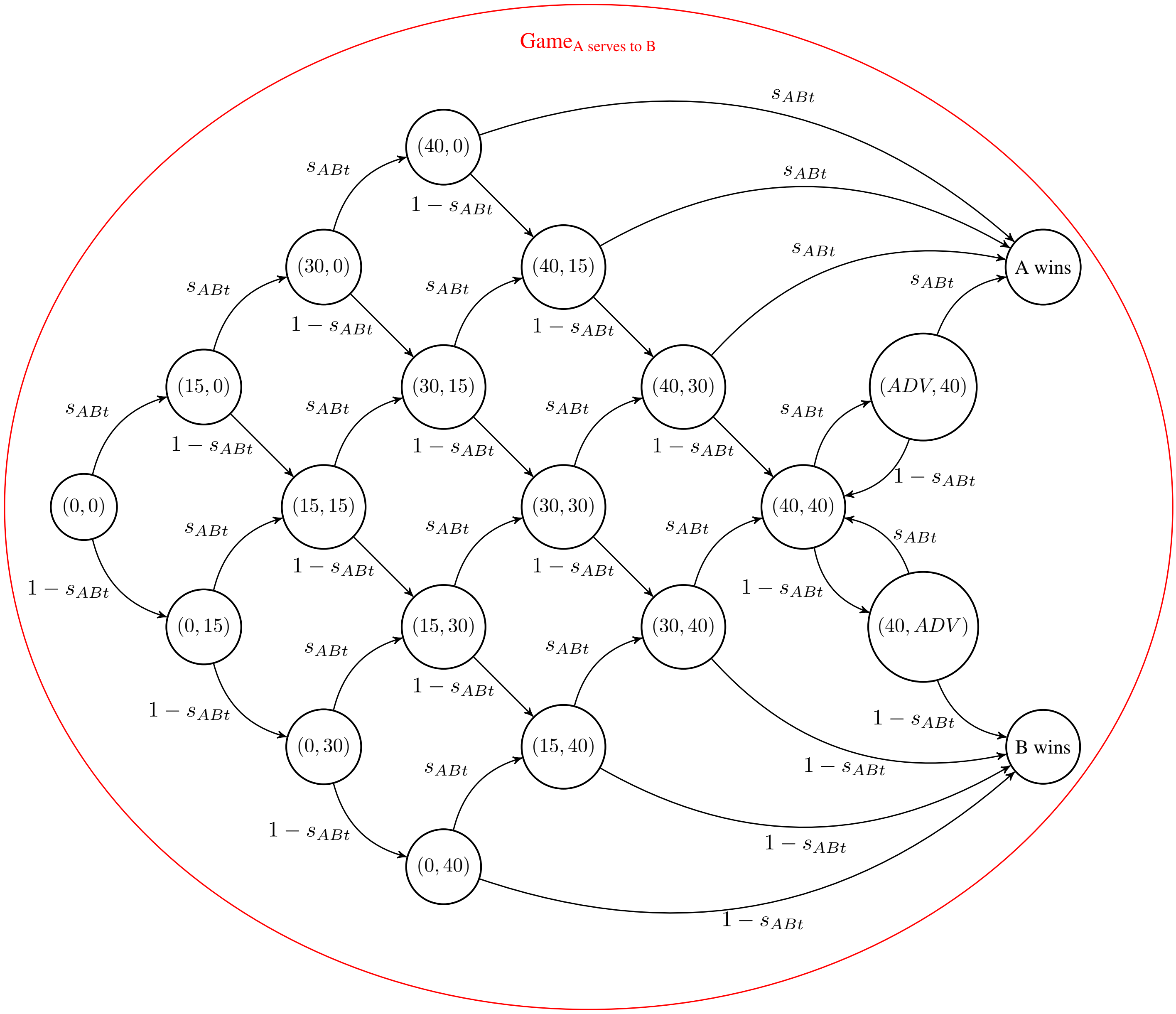
Modeling the Tennis Scoring System as a Markov Chain

Consider a match between Player A and Player B. At any point in the match, we can read the score of the match as (number of sets A has won, number of games A has won in the current set, number of points A has won in the current game, number of sets B has won, number of games B has won in the current set, number of points B has won in the current game, current server). We can take the current score as states in a chain and define transition probabilities as the probability that A (or B) wins the next point, game, tiebreak, or set. I provide separate Markov chains at the game, set, and match levels due to visualization limitations but I think it's fairly intuitive how these chains can be combined, while carefully keeping track of the current server, to properly model a full tennis match.

Game-view Markov Chain

At match t, say Player A wins a serve against Player B with probability s_{ABt} and Player B wins a serve against Player A with probability s_{BAt}, where each serve is independent and identically distributed (iid)[1] given the server. In a game where A serves to B, I create a graph in which nodes represent a possible score attained in the game written as (A's score, B's score). There are two terminal nodes, "A wins" or "B wins" since either one player or the other wins the game. With our assumption that each time A serves to B, A wins with probability s_{ABt}, where each serve is iid, I can label transition probabilities between nodes as s_{ABt} or $1 - s_{ABt}$ depending on if the transition constitutes A winning a point or B. I can then numerically compute the probability that we reach node "A wins" (or node "B wins") from node (0,0) and use that as our estimate that A (or B wins the game).

[1] The Markov assumption is that transition probabilities be depend only on the current state (not on the past), transition probabilities do not need to be iid given the server as I've imposed. I make this assumption for simplicity of the model and implementation.

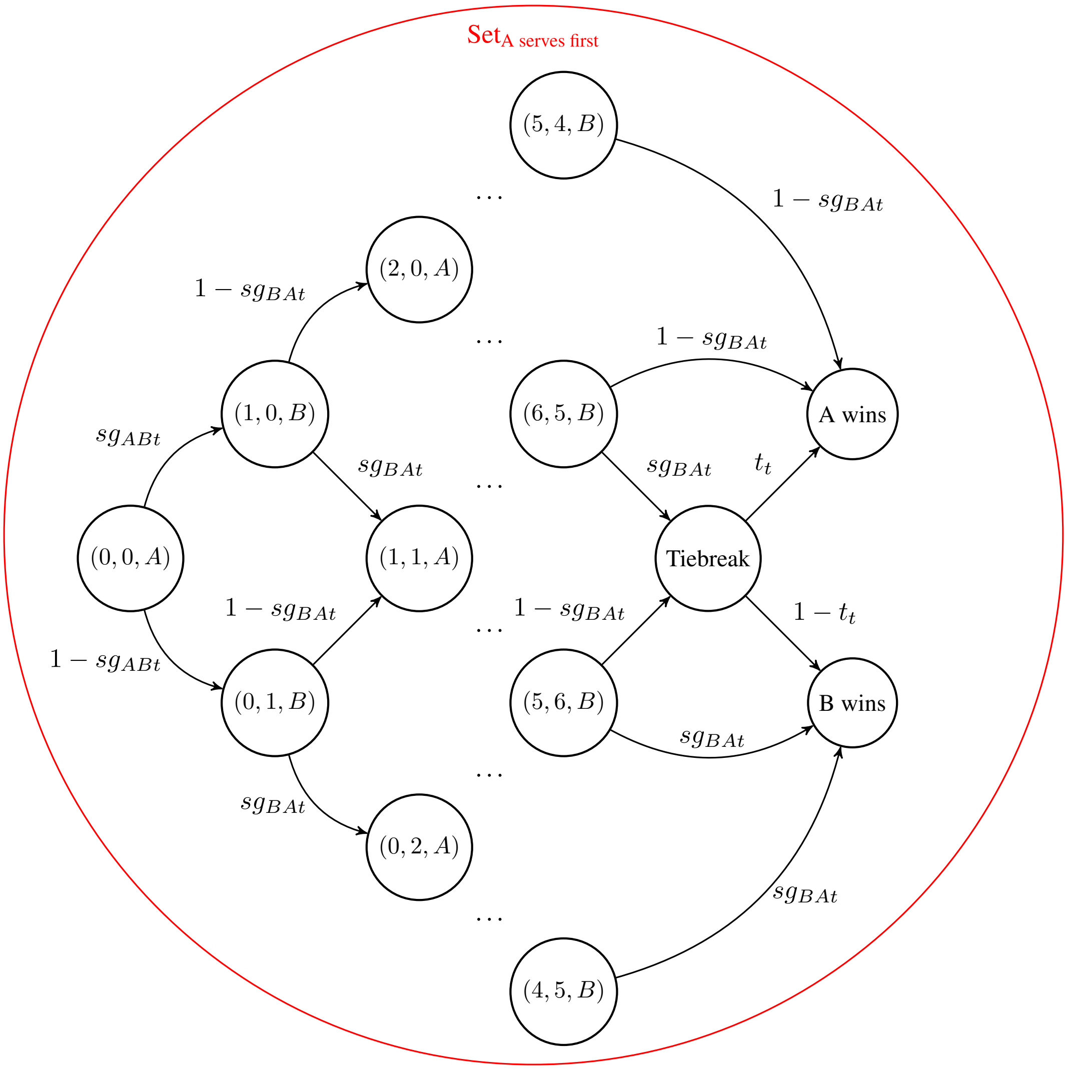


Set-view Markov Chain

Similar to the game level, I can construct a Markov chain for the set. I create nodes that represent possible scores in the set: (num of games A has won in the set, num of games B has won in the set, server). I add in transition probabilities that reflect when A wins the game or B wins and I add in two terminal nodes ”A wins” or ”B wins”, that are states where A has won the set or B. At match t, say Player A wins a service game against Player B with probability sg_{ABt} and Player B wins a service game against Player A with probability sg_{BAt}, where each service game is independent and identically distributed (iid) with its respective probability depending on the server[2]. If I choose to discretize the game at the point level, I can use the computed probability in Game$_{\text{A serves to B}}$ that A wins the game to assign transition probabilities from states where A serves, vis-a-vis for Player B. Alternatively, I can discretize the set at the game level and take sg_{ABt} and sg_{BAt} as parameters to the model. Lastly, I need to take special care for the tiebreak at the end of the set. At match t, say Player A wins a tiebreak against Player B with probability t_t, where each tiebreak is independent and identically distributed (iid). I can take t_t as a native parameter to the model or use s_{ABt} and s_{BAt} to numerically compute t_t. Finally, given my transition probabilities, I can solve for the probability that each player wins the set by finding the probability of

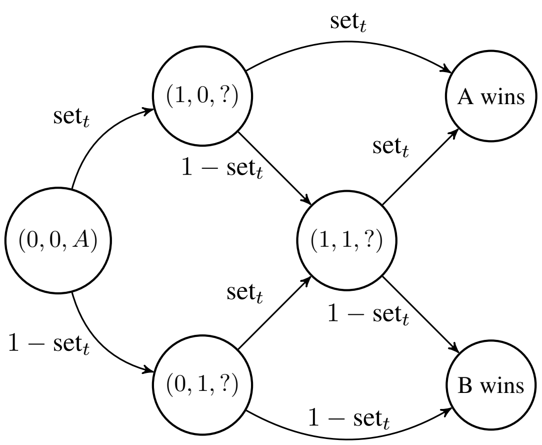
[2] Again, I impose this iid assumption to simplify the model and its implementation.

reaching each terminal state from $(0, 0, A)$.



Match-view Markov Chain

Lastly, I can also construct an analogous Markov chain for the match where nodes are the score in sets: (Num of sets A has won, Num of sets B has won, first server of set). At match t, say Player A wins a set against Player B with probability set$_t$, where each set is independent and identically distributed (iid). I add in transition probabilities according to the probability that each player wins a set. I also add in two terminal nodes, "A wins" and "B wins" that indicate Player A or B has won the match, respectively. Similar to above, set$_t$ can be made an exogenous parameter to the model or we can compute it using the Set$_{\text{A serves first}}$, Set$_{\text{B serves first}}$ Markov chains. I draw the relatively small Markov chain for a best-of-three-set match. The question marks for the servers reflect that at this granularity, we don't know the final server in the each set and so we don't know who serves first in the sets following the first. While a problem here, this issue doesn't arise in a full model of the tennis match since we would just have separate states for either server and add proper transitions.



Costs and Benefits of the Markov Model

Since the sport of tennis progresses in discrete points, games, and sets, I can fairly easily model a match as a state-based system where states are possible scores in the match. In constructing a Markov model from these states, we inherently undertake the Markov assumption: we assume that that all transitions probabilities from a state depend only on the current state. I take a further assumption in my analysis than the Markov Model requisites for simplicity. I assume that all points served in a match by a server are identically distributed. Are these good assumptions? Likely no. In a match between A and B, even if A has won the past 10 points, I assume that doesn't affect the likelihood they win the next point. The "hot-hand principle", the occurrences of streaks in "random" sequences, has been fervently debated in papers such as Gilovich, Vallone, and Tversky (1985), arguing that it is a fallacy and Sanjurju and Miller (2018), in a beautiful result, arguing that GVT's selection of streaks to examine is biased, invalidating their work. In the lens of tennis, Klaassen and Magnus (2001a) finds that tennis points are not independent and identically distributed[3]. For example, key points in the match such as "match point", are less likely to be won by the server than a typical point. Nevertheless, they find that deviations from iid points are typically small.

A simple alternative to my further assumption (that still satisfies the Markov assumption) is to make the serve win probability indexed to the score in the current game like in Matteazzi and Lisi (2017). For example, we could define $s_{ABt(0,0)}$, $s_{ABt(15,0)}$, etc., as the probability that at match t, player A wins a service point on player B when the score is (0,0) in the current game, (15,0) in the current game, respectively, etc. Such indexed service point win probabilities would also partly help control for streaks since a score of $(40, 0)$ necessarily means that the server won the past three points. There are many other ways to define states of the tennis match to take into account the past few points, games, and sets to generalize the iid points assumption- these extensions are out of the scope of this article.

There are a few more nice applications of the Markov model. First, as discussed in Klaassen and Magnus (2001b), we can efficiently solve for the probability of winning the match at any given state by analytically solving any cycles in the chain (ie., deuce), and then using dynamic programming to find the win probability from any state. Discretizing the match at the point level and given serve win probabilities s_{ABt} and s_{BAt}, and scoring rules for the match (ie., number of sets needed to win the match, etc.), a memoized function takes the following inputs.

```
# Global parameters
s_ABt = 0.8 # probability A wins a service point on B
s_BAt = 0.75 # probability B wins a service point on A
sets_to_win = 2 # number of sets a player needs to win to win the match
games_to_win = 6 # number of games a player needs to win to win the set
points_to_win = 4 # number of points a player needs to win to win the game
tiebreak_total = 7 # number of points a player needs to win to win a tiebreak

def compute_playerA_win_prbability(A_sets, B_sets, A_games, B_games, A_points, B_points, server):
```

[3] They conclude that winning the previous point has a positive effect on winning the current point. Though, since the authors wrote their paper in 2001 and the streak selection bias found by SM happened in 2018, KM's paper doesn't account for this bias. As explained in the appendix, this bias causes one to underestimate the presence of the "hot-hand" phenomenon. Thus, SM's implication only exacerbates KM's finding that a player winning sequential points are correlated events

```
    """ Return the probability that A wins the match from the current score and global parameters.

    Args:
    - A_sets: number of sets A has won
    - A_games: number of games A has won in the current set
    - A_points: number of points A has won in the current game/ tiebreak
    - B_sets: number of sets B has won
    - B_games: number of games B has won in the current set
    - B_points: number of points B has won in the current game
    - server: the server of the next point

    Return Value:
    - int in [0,1]: the probability A wins the match from the current score
    """
```

Also explained in this KM (2001b) paper, the Markov chain approach allows us to ponder changes to the tennis scoring rules (ie., sets are first to 8 games, no tiebreak) and how they impact the win probability of the match. Lastly as discussed in Gollub (2017), the Markov model allows us to compute live in-match win probabilities by computing the match win probability for a player from any score reached in the match.

REFERENCES

Thomas Gilovich, Robert Vallone, and Amos Tversky. The Hot Hand in Basketball: On the Misperception of Random Sequences. *Cognitive Psychology*, 17(3):295–314, 1985. URL: `https://www.sciencedirect.com/science/article/pii/0010028585900106`.

Jacob Gollub. Producing Win Probabilities for Professional Tennis Matches from any Score. *Bachelor's thesis, Harvard College*, 2017. URL: `http://nrs.harvard.edu/urn-3:HUL.InstRepos:41024787`.

Franc J G M Klaassen and Jan R Magnus. Are Points in Tennis Independent and Identically Distributed? Evidence From a Dynamic Binary Panel Data Model. *Journal of the American Statistical Association*, 96(454):500–509, 2001. URL: `https://doi.org/10.1198/016214501753168217`.

Francesco Matteazzi and Francesco Lisi. A follow-up study on the issue of i.i.d. points in tennis, 2017. URL: `http://www.mathsportinternational2017.math.unipd.it/slides/MS2017_Matteazzi.pdf`.

Joshua B. Miller and Adam Sanjurjo. Surprised by the Hot Hand Fallacy? A Truth in the Law of Small Numbers. *Econometrica*, 86(6):2019–2047, 2018. URL: `http://www.jstor.org/stable/44955325`.