ATP and WTA Professional Tennis: Calculating In-Match-Win Probability with Bayesian Modeling

Ben Moolman

Table of contents

0.1 Introduction

In this project, we explore the in-match-win probability of professional tennis matches. Tennis' scoring format allows for huge momentum swings in a short amount of time. We are going to explore how the probabilities that tennis players win a match are calculated and update throughout the match using Bayesian modeling. To determine the probability that a player (player 1) wins a match against another player (player 2), we incorporate (1) the probability that player 1 wins a point on serve against player 2, (2) the probability that player 2 wins a

point on serve against player 1, and (3) the current score of the match of interest. The prior distributions for (1) and (2) are generated from points played in matches prior to the match of interest. Both (1) and (2) are then updated throughout the match of interest as points played between the two players update their prior point-win probabilities. As case studies, we explore the 2022 Men's US Open Quarterfinal between Carlos Alcaraz and Jannik Sinner and the 2023 Women's US Open Final between Coco Gauff and Arnya Sabalenka.

0.1.1 Outline

We begin by exploring tennis scoring to get an understanding of how a match is played in Section **??**. We then look at the data we are using in this project on Section **??**. Then, in Section **??**, we discuss Bayesian modeling, and come up with our probabilities of the players winning a point on their serves against specific opponents. We calculate the in-match-win probability and how we update our prior distributions throughout the match. We then look at the case studies of the 2022 Men's US Open Quarterfinal between Carlos Alcaraz and Jannik Sinner, and discuss how we can change what matches we include in our prior distribution in Section **??** and Section **??**. We also explore the 2023 Women's US Open Final between Coco Gauff and Aryna Sabalenka in Section **??**. We conclude in Section **??** with a discussion of the results and potential future work. An appendix in Section **??** is included with the code used in this project for reference.

0.1.2 Tennis Scoring

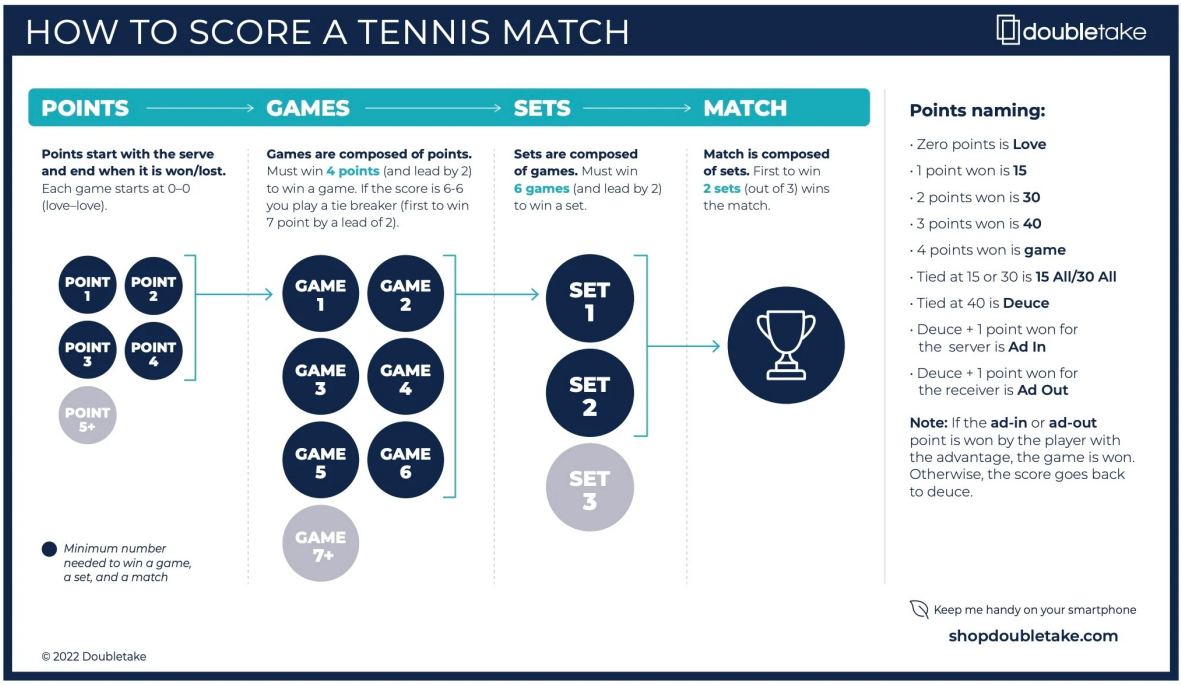
The scoring format in tennis can be confusing to those that are not familiar with the sport. Below is a brief outline of how tennis matches are scored:

- The match starts with one player serving at 0-0
- The server is the player that hits the ball first in a point
- The server alternates what direction they serve from after every point (deuce side and ad side)
- A game is played with the server serving for the entire game
- A game is won by the first player to win 4 points, wining by a margin of 2 or more points
- Sets are played first to 6 games, wining by a margin of 2 or more games
- If the score reaches 6-6 in a set, a tiebreak is played
- A tiebreak is won by the first player to win 7 points, wining by a margin of 2 or more points
- A match is played to the best of 3 or 5 sets based on the tournament format

A little more tennis terminology are the words break and hold, defined below:

- Break: when the player returning wins the game
- Hold: when the player serving wins the game

A graphic is attached below to help, and a more detailed write-up can be found on the website *doubletake* here.



0.1.3 Data

The data used in this project is from the ATP and WTA professional tennis tours, and is from Jeff Sackman's tennis data on Github. There is point-level data on the ATP and WTA main-draw singles grand slam tournaments from 2011-present. There is also match-level data for ATP matches and match-level data for WTA matches. Access to the data is needed to ensure correct spelling of player names (especially nicknames, ie "Rafa Nadal" vs "Rafael Nadal") and tournament names, the correct date ranges, file paths, and match IDs for specific matches of interest. Checking these details is best done on the matches data frames, and if finding the information on the Github site is problematic, the data can be read in using the `read_matches()` function, which works for both ATP and WTA data.