# 2024美赛论文初稿

## 摘要

## 问题一模型的建立与求解

### Content of the problem

Develop a model that captures the flow of play as points occur and apply it to one or more of the matches. Your model should identify which player is performing better at a given time in the match, as well as how much better they are performing. Provide a visualization based on your model to depict the match flow. Note: in tennis, the player serving has a much higher probability of winning the point/game. You may wish to factor this into your model in some way.

### Analysis of the problem

The problem requires us to develop a model that dynamically captures the performance status of both players in each points of a tennis match and presents it in a quantitative form to show the performance status of the players. At the same time, we hope that this model is universally applicable to valuable tennis match scenarios. Therefore, we consider starting from the scoring points, based on which we explore and consider multiple factors, in order to establish a comprehensive mathematical model to quantify the performance of the players; on the other hand, for the optimal selection of some parameters in the model, we consider using the whole dataset given in the title to carry out training, and according to the results of the training to guide the calibration of the relevant parameters, which in fact forms a model tuning BP process.

### Modeling

We believe that there exists a "momentum" of the athlete's own state, which reflects the athlete's ability to master the game, the degree of his or her own strength, and the ability to reverse an unfavorable situation in the course of the game. This likelihood is related to the influence of the events that the athlete is subject to, such as whether the point is scored or not, whether he/she has the right to serve or not, etc., and it has a significant positive correlation with his/her probability of scoring a point.

Furthermore, we build a winning performance model based on the conditions and flow of the matches, which determines whether a player is performing well or not based only on the score-dependent game state parameter $X$ (note that we do not take into account factors that are closely related to the player's own characteristics, such as stamina level, resistance to stress, etc., in order to ensure the universality of the model), which consists of the following parameters:

Right to serve: $x_r : \{1, 0\}$, 1 means the right belongs to you, 0 means the right belongs to the opponent.

Points scored/missed: $N : \{1, 0\}$, 1 indicates a point scored, 0 indicates a point scored by the opponent.

Number of consecutive points scored/missed: $x_1$, positive integer, indicates the number of consecutive points scored or missed.

Whether or not the serving team made an error: $x_2 : \{0, 1\}$, 0 means no serving error, 1 means serving error.

Whether it is a game point or not: $x_g : \{1, 0\}$, 1 means it is a game point, 0 means it is not.

Whether it is a set point: $x_s : \{1, 0\}$, 1 means it is a set point, 0 means it is not.

Whether it is a direct service point: $x_3 : \{1, 0\}$, 1 means it is a direct service point, 0 means it is not.

ariable $x$ above is time-dependent $x(t_n)$ and the state parameter $X(t_n) = (x_r, N, x_1, x_2, x_3, x_4, x_g, x_s)$

Define the momentum metric function $Q(t_n) = kQ(t_{n-1}) + F(X(t_{n-1}))$, where$n = 1, 2, 3 \cdots,\ Q(t_1) = 0$, $t_1$ denotes the time of the first scoring point, and the decay rate, $k \in (0, 1)$, is is used to measure the extent to which varying time away from that score point affects the momentum of the next scoring point, and we believe that the closer the event is to that scoring point, the more it will play a role in influencing the momentum of the next scoring point.

At the same time we define the predicted win rate:

$$P = \frac{Q_1(t_n) - Q_2(t_n)}{Q_1(t_n) + Q_2(t_n)}$$

Representing the predicted win rate for that score point derived from the quantized momentum, we argue that the greater the difference in momentum between player one and player two when player one's momentum is higher than player two's at that score point, the greater the likelihood that player one should be victorious at that score point.

Under this definition, F denotes the upward tendency of the player's winning percentage at the n-1st score point due to various game performances, and is used by us to evaluate the player's performance. The performance evaluation function is defined as follows:

$$F(X(t_{n-1})) = [(1 - x_r)\frac{q}{1-q} + x_r]\delta(N, 1)[1 + k_1(x_1 - 1) - k_2 x_r x_2 + k_3(x_g + x_s)]$$
$$-k_4[\frac{q}{1-q}x_r + (1 - x_r)]\delta(N, 0)[1 + k_1(x_1 - 1) + k_5 x_2 x_r + k_3(x_g + x_s) + k_6 x_3]$$

Among that we consider:

1. the right to serve gives the player a greater advantage, we counted the average winning percentage q of the serving side in the dataset, and used the reward coefficient $\frac{q}{1-q}$ to amplify the performance score when the receiving side wins and the performance deduction when the serving side loses, here $q \approx 0.6731$;
2. when scoring $N = 1$, $\delta(N, 1) = 1, \delta(N, 0) = 0$, only the first term of F is non-zero, at this time, if the other side serves the ball, you get the reward coefficient $\frac{q}{1-q}$, otherwise, you get 1, and the same when $N = 0$. That is, the first term of F is the positive rating when winning the game, and the second term is the negative rating when losing the game;
3. The last part of the two terms of F determines the base value of the evaluation score, including the win/loss situation, consecutive points scored or lost, and the effect of errors on one's own serve. In addition, wins and losses at set and game points have a greater impact on the player than those in general;

4. the scoring factors are related to each other, which in order to incorporate the examination of the model, such as the situation of the ball possession and the situation of the stronghold can reflect whether it is a break point or not.

On this basis, we define the optimization objective as

$$min\left(T=\frac{1}{N_{play}}\sum_{all}\sum_{t_n}\left(\frac{Q_1(t_n)-Q_2(t_n)}{Q_1(t_n)+Q_2(t_n)}-N(t_n)\right)^2\right)$$

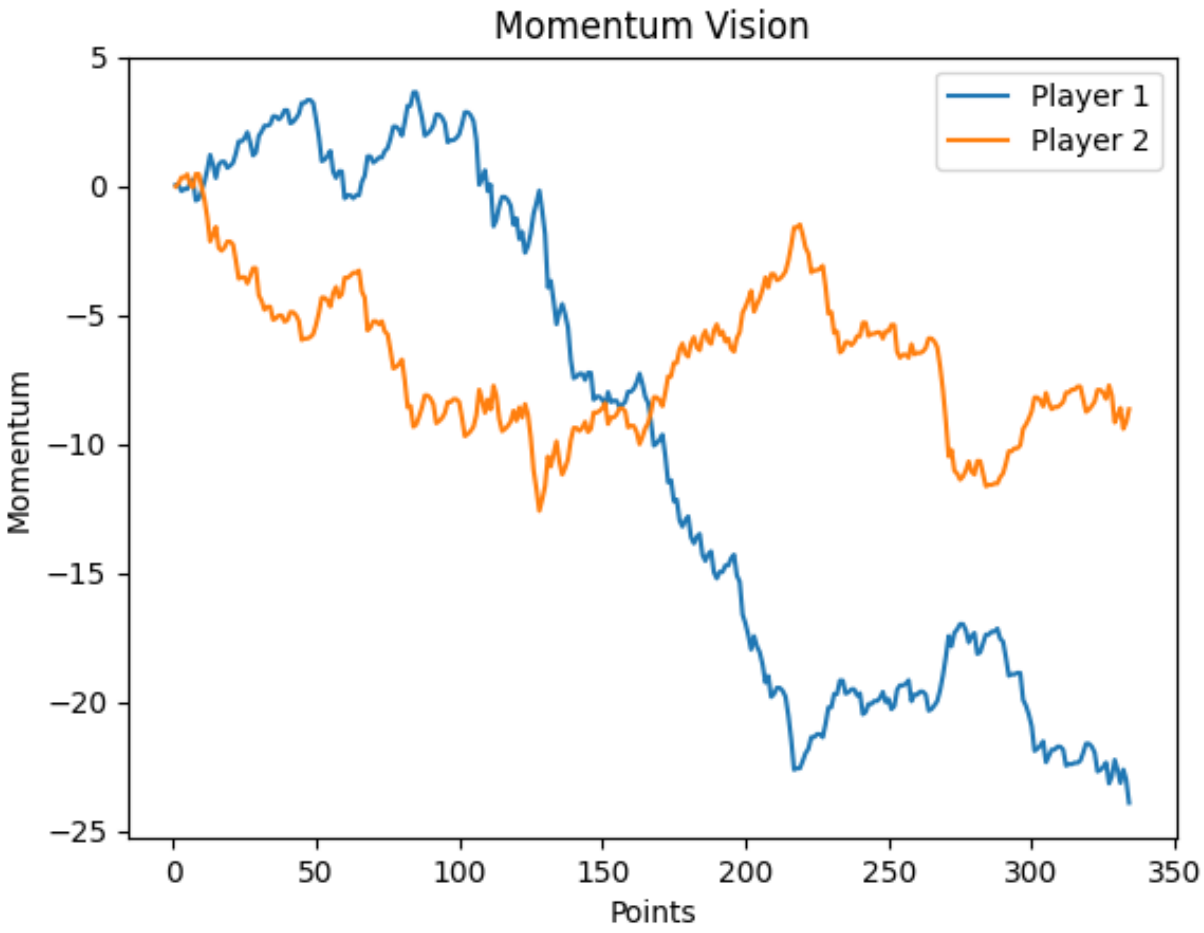
The optimization goal is to make the predicted wins as close as possible to the actual wins. We include every scoring point of all games in the dataset in the training dataset, and parameterize it to obtain a universal evaluation model.

## Solving and Analysis

First we processed and extracted the dataset provided by the question using python code, divided it into thirty-one complete matches, and extracted the required factors from each match, such as the right to serve, the number of serves, and so on, which in turn were transformed into the factors required for the performance evaluation function, such as consecutive points scored or lost, whether it was a set point or not, and so on. In turn, we achieve a complete reproduction of the above model and construct a functional form of the optimization objective, which is determined by six parameters, and whose function return value we expect to be as small as possible.

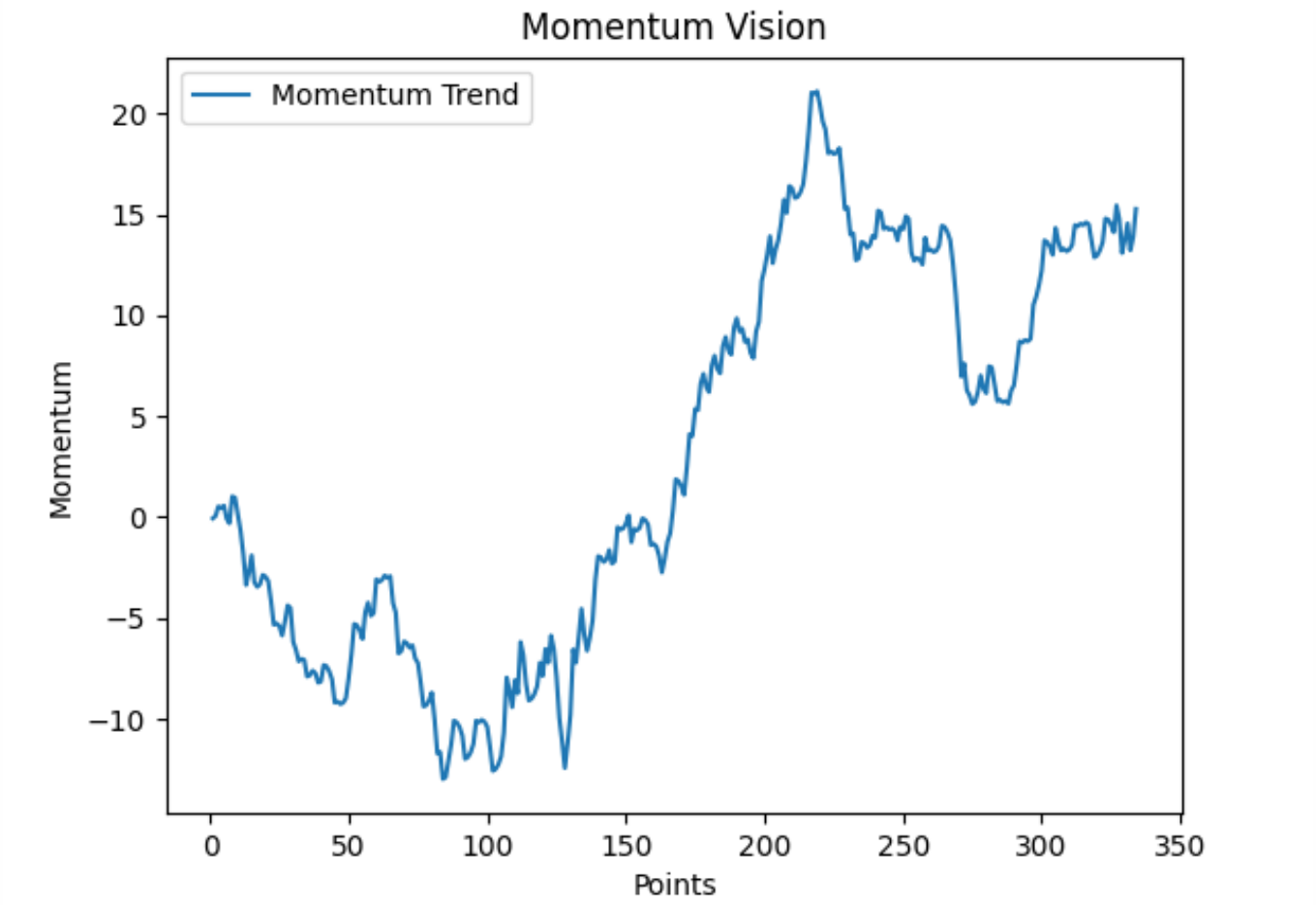
Using the Gradient Descent Method, the parameters were iteratively adjusted by BP, and the above objective function value was finally optimized to 0.4992, which indicates that the momentum predicted win rate conforms to the actual situation to approximately 0.65.

At the same time, we obtain a near-optimal solution for all k-parameters in the F function:
$\vec{k}=(56.46,-26.96,195.48,-190.65,28.36,49.67)$. Taking the final match played by Carlos Alcaraz against Novak Djokovic as an example, we plotted images of the variation of momentum $Q_1(t_n)$ and $Q_2(t_n)$ for both:



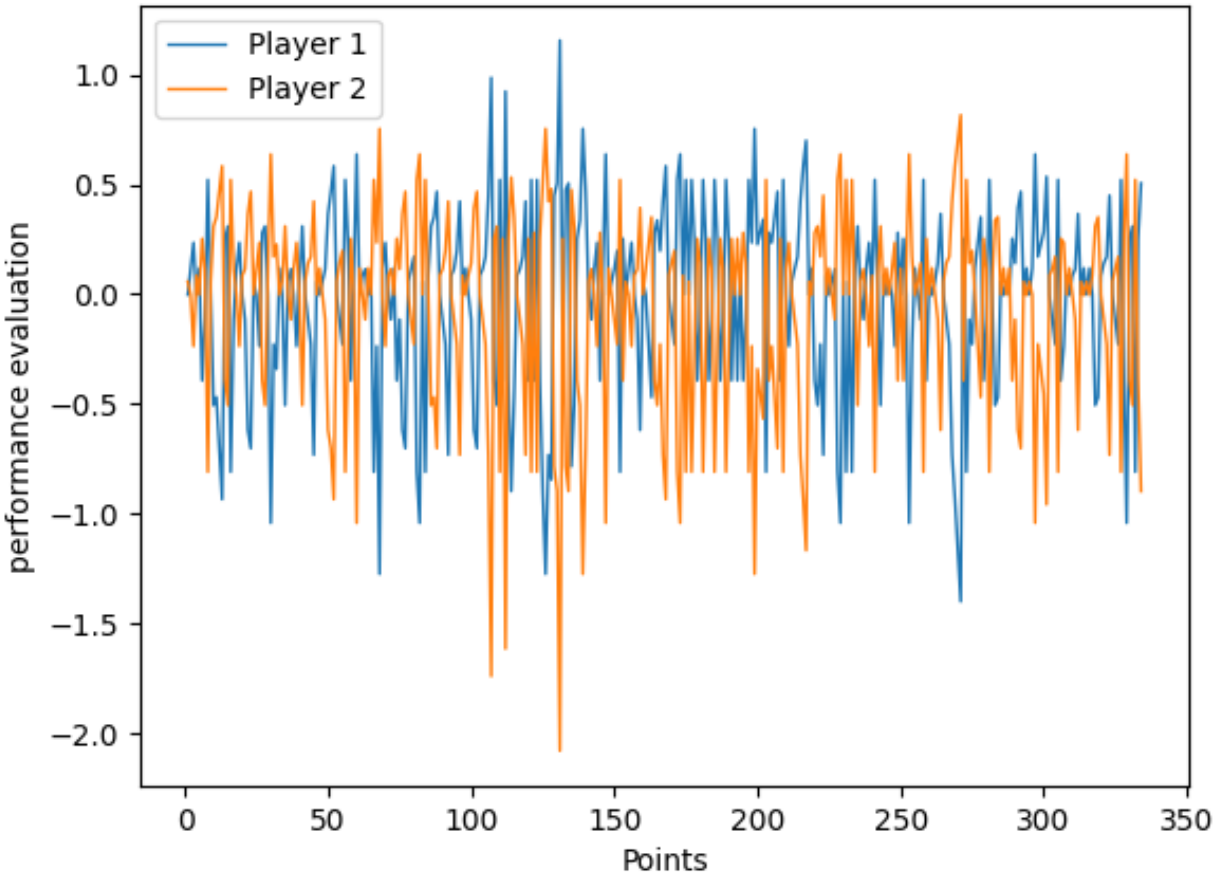

As you can see from the graph (where player 1 is Djokovic and player 2 is Alcarez), when the match first started, Djokovic was firmly in control of the match and thus dominated most of the points in the first set, however, as the match progressed, the momentum of the two athletes began to change, which allowed Alcarez started to fight back in the second set and took a hard fought set in a tie-breaker. Thereafter, Djokovic began to gradually take control of the match and won the third set comfortably with a huge 5 game difference. Although Djokovic regained a set in the fourth set (which is reflected in a small peak in the blue line in the graph late in the match), the match ultimately tipped the scales in favor of Alcarez, which is a reflection of Alcarez's momentum.

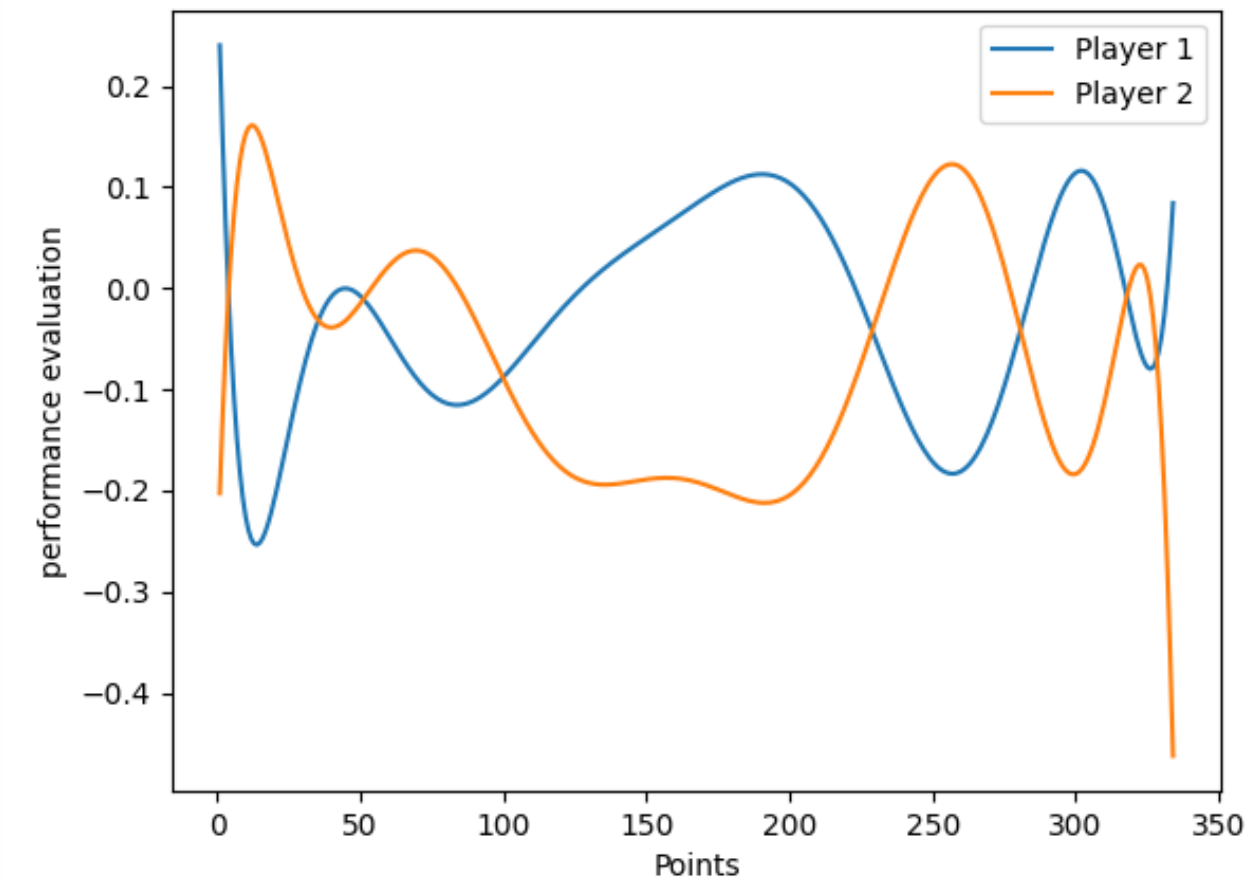
At the same time, we visualized the difference between the momentum of the two athletes in the race:




This image provides another, more intuitive perspective on the flow of the match, as we can easily find which player is performing better at a given moment in time, which is reflected in the plus and minus of the vertical axis, while the absolute value of the vertical axis value quantifies how well the player is performing. In addition, looking at the fluctuations of the curve, it is easy to see that some of the turning points of the curve reflect the tendency that players are trying to take the pace of the game into their own hands.

In addition, we plotted the image of the performance evaluation function $F(X(t_{n-1}))$ of the two players in the flow of the game, and at the same time, in order to better observe the trend of its value through the more violent fluctuations, we made a certain fit to the discrete sequence composed of the function value, in order to fully show its relative trend on the basis of filtering out the violent fluctuations, as shown in the figure:


Player 1
Player 2
performance evaluation
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
0
50
100
150
200
250
300
350
Points




The curve formed by the values returned by the performance evaluation function can be seen in the graph, and its general trend and relative magnitude are similar to the course of the game analyzed earlier, but what can be observed in particular is that towards the end of the game, the value of Djokovic's performance evaluation begins to drop drastically, which is considered to be a sign that, on one hand, he has completely lost the control over the situation, and, on the other hand, it also reveals that there may be the influence of some other potential factors in the course of the process, and that's exactly what we will need to take into account later on.

All in all, these charts give us a visual representation of the flow of the game and the performances of both players, and provide different perspectives for us to cut through and get a fuller picture of the game.

# 问题二模型的建立与求解

## 问题内容

A tennis coach is skeptical that “momentum” plays any role in the match. Instead, he postulates that swings in play and runs of success by one player are random. Use your model/metric to assess this claim.

## 问题分析

The tennis coach 认为比赛过程中的swings是随机的，不依赖于所谓势能。依照我们的定义，momentum是由过去一系列比赛进程及状况积累起来的、对当前得分点是否得分的影响，这种影响就体现在每一个得分点的swings中。也就是说，教练认为，当前得分点事件发生的结果只依赖于当前比赛条件，而与前面得分点的相关事件无关，即认为网球比赛的赛况是一个马尔科夫链的过程；而如果认为比赛过程中存在momentum，即该过程是一个非马尔可夫链的过程。因此，为了判断网球赛程是否为马尔可夫链，我们在第一问模型的基础上，将模型中非马尔科夫的项$Q(t_{n-1})$删去，把表现函数转化为当下表现函数，将状态参量$X(t_n)$替换为$X'(t_n)$后，进行相同的训练，得到代表数据的马尔可夫链模型，并与原模型进行比较以得出最终结果。

## 模型建立

定义状态空间$S=\{(0,0,0),(1,0,0),\dots\}$，表示当前的成绩（points、games、sets），用$Y(t_n)$表示$t_n$时刻的成绩，则基于马尔可夫链的模型满足：

$$\begin{aligned}P\{Y(t_n)=s(t_n)|Y(t_1)=s(t_1),Y(t_2)=s(t_2),\dots,Y(t_{n-1})=s(t_{n-1})\}\\=P\{Y(t_n)=s(t_n)|Y(t_{n-1})=s(t_{n-1})\}\end{aligned}$$

$$P\{Y(t_n)=s(t_n)\}=F_{unknown}(X'(t_n),Y(t_{n-1}))$$

其中$X'(t_n)=(N(t_{n-1}),x_1(t_{n-1}),x_3(t_{n-1}),x_r(t_n),x_g(t_n),x_s(t_n),x_2=0)$表示马尔科夫链中的当下状态。

为了保持两种模型形式上的一致，便于后续进行两种模型的比较，我们选取了第一问模型中主要的状态参量，但采用了不同得分点的构成。我们认为，当下得分点的状态由两部分组成，一是上一点的得分情况（胜者/连续得分次数/发球直接得分），二是当前得分点未得分时就有的性质（局点/盘点/球权）。

令$Q'(t_n)=F(X'(t_n))$，用当下表现函数代替第一问模型中的表现函数，则获胜概率可以表示为（其中概率为负表示对方胜率）：

$$P\{Y(t_n)=s(t_n)=Y(t_{n-1})+(1,0,0)\}=\frac{F(X_1'(t_n))-F(X_2'(t_n))}{F(X_1'(t_n))+F(X_2'(t_n))}$$

得到与第一问形式相似的优化目标：

$$min\left(T'=\frac{1}{N_{play}}\sum_{all}\sum_{t_n}\left(\frac{F(X_1'(t_n))-F(X_2'(t_n))}{F(X_1'(t_n))+F(X_2'(t_n))}-N(t_n)\right)^2\right)$$

对训练得到的马尔科夫链模型的表现函数$F(X'(t_n))$，我们将其优化目标的收敛值、收敛精度与第一问中的结果进行对比。此外，我们比较两种模型逐点预测真实序列的准确度，即从数据集中任取一场比赛的比赛结果作为真实数据，其真实胜负结果N作为数据序列，也即$N(t_n)$，逐点将得分点信息输入最优化后的Q模型和F模型，根据相应的概率公式输出预测的下一得分点的胜负概率，得到一串胜负概率序列$Q\{P_{t_n}\}$及$F\{P_{t_n}\}$，分别计算并比较两者大小：

$$\sum_{t_n}(q_n-N(t_n))^2$$

$$\sum_{t_n}(f_n-N(t_n))^2$$

通过对两种模型的多维度综合对比，以及结合真实值的比较，可以充分地判断出网球赛程是否为马尔可夫链。

## 模型求解与结果分析

对该目标进行优化得到最优$T' =$，其优化精度$\delta T' =$，以及最优参数$\vec{k} = (k_1, k_2, k_3, k_4, k_5, k_6)$

这样，我们就得到了马尔科夫链模型的表现函数$F(X'(t_n))$。

第一问中，优化目标的收敛值为$T =$，收敛精度为$\delta T$。可以发现，与第一问的优化结果相比，$T'$显著大于$T$，且$\frac{T'-T}{\delta T} =? >> 1$，因此不考虑前置比赛条件的当下表现模型对胜率的最优预测显著劣于考虑了前置比赛条件的momentum-表现模型。

# 问题三模型的建立与求解

## The First Problem内容

Coaches would love to know if there are indicators that can help determine when the flow of play is about to change from favoring one player to the other.

Using the data provided for at least one match, develop a model that predicts these swings in the match. What factors seem most related (if any)?

## The Second Problems内容

Given the differential in past match “momentum” swings how do you advise a player going into a new match against a different player?

（此得分点球员体力$H(i,t)$

其中体力依赖于：

前一时刻的体力、此得分点的跑动距离、此得分点的击球次数、此得分点的击球速度

（此外，考虑特殊得分点，如抢七抢十。）

考虑哪些因素影响势能的转换，并预测势能的转换

前面我们已经成功模拟了比赛流程中的势能量化，其中$\Delta Q = 0$时即发生势能转换，此处注意，势能转换具有方向性。

此处不难认为，加权系数大的因素，对势能的波动更具有影响力。

前面我们是在已知该得分点的全部数据下，计算出该得分点时的球员势能，现在要求我们对此势能波动进行预测，此处考虑是仅对势能转换点进行预测，还是全部波动进行预测。

前者为二分类问题，可以采用随机森林、微分方程方法

后者难度较大，考虑预测后与真实值进行校对后进行下一阶段的预测）

# 问题四模型的建立与求解

## 问题内容

Test the model you developed on one or more of the other matches. How well do you predict the swings in the match? If the model performs poorly at times, can you identify any factors that might need to be included in future models? How generalizable is your model to other matches (such as Women's matches), tournaments, court surfaces, and other sports such as table tennis.

## 问题分析

## 模型建立

## 模型求解与结果分析

（建立检验标准来量化预测准确程度

进行迁移学习来检验泛化能力

考虑不同数据集应当更改、添加和删除的关键影响因素）