Coursework 2 - Probabilistic rankings

Candidate number: H801L
Word count: 1000

19th November 2020

1 Part A

1.1 Code

We used Equation 1 to compute the likelihood mean, which is the sum of all game performances for a player. Equations 2 and 3 were used to compute likelihood precision matrix, which is related to how many games each player played in total (diagonal) and with specific players (off-diagonal).

$$\tilde{\mu}_i = \sum_{g=1}^{G} t_g \delta\left(i - I_g\right) - \sum_{g=1}^{G} t_g \delta\left(i - J_g\right) \tag{1}$$

$$\left[\tilde{\Sigma}^{-1}\right]_{ii} = \sum_{g=1}^{G} \delta\left(i - I_g\right) + \delta\left(i - J_g\right) \tag{2}$$

$$\left[\tilde{\Sigma}^{-1}\right]_{i \neq j} = -\sum_{g=1}^{G} \delta\left(i - I_g\right)\delta\left(j - J_g\right) + \delta\left(i - J_g\right)\delta\left(j - I_g\right) \tag{3}$$

```
m = np.zeros((M, 1))
for p in range(M):
    # TODO: Compute likelihood mean term
    m[p] =  np.sum(t[G[:,0]==p]) - np.sum(t[G[:,1]==p])
iS = np.zeros((M, M))

for g in range(N):
    # TODO: Compute precision matrix
    iS[G[g,0], G[g,0]] += 1
    iS[G[g,1], G[g,1]] += 1
    iS[G[g,0], G[g,1]] -= 1
    iS[G[g,1], G[g,0]] -= 1
```

Listing 1: Sample of gibbsrank.py with implemented code

1.2 Problem Analysis

It is hard to determine **Burn-in** time for Gibbs sampling, by simply considering skill levels of players, Figure 1. To investigate properties of the underlying distribution we perform Monte Carlo simulation of 50 Gibbs sampling runs, Figures 2 and 3.

We look at skills of the top player, Djokovic, because we expect skills of best players to experience **worst initialisation** and **lowest mixing-rate**. This is because their posterior skills will lie away from the prior $p(w) \sim N(0, 0.5)$, and their skill variance is lower than average, due to many games played.

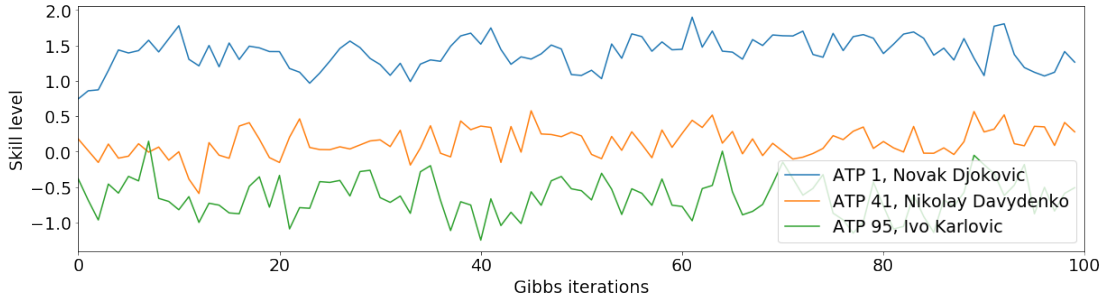


Figure 1: Skill samples for three players at different iterations of Gibbs sampling.

1.3 Monte Carlo simulation

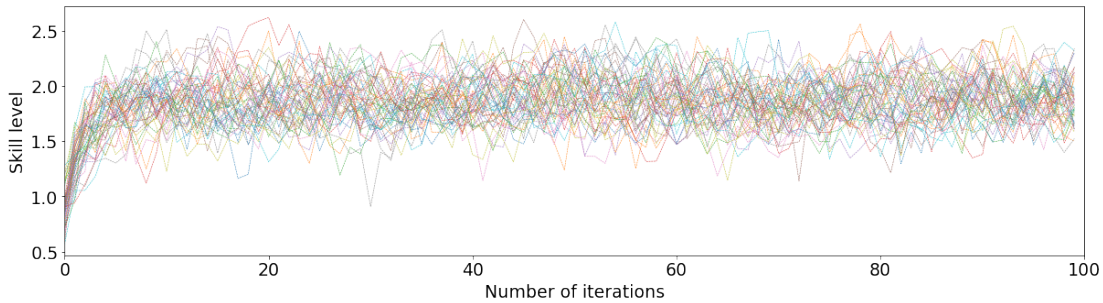


Figure 2: 50 trials of Monte Carlo simulation for Novak Djokovic skills.

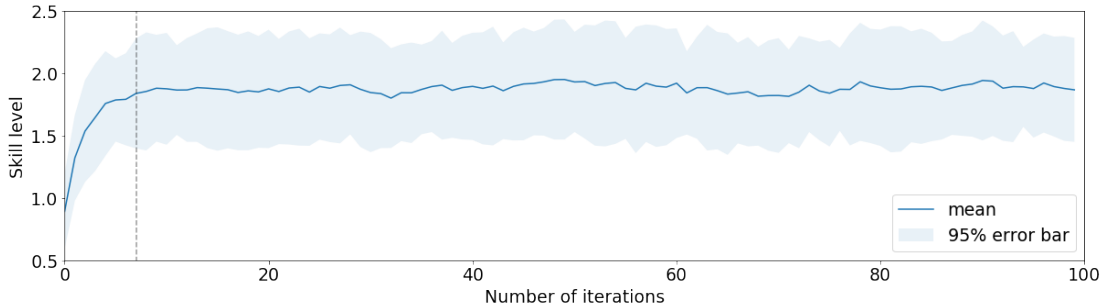


Figure 3: Mean and 95% confidence bar for 50 trials of MC in Figure 2. Convergence seen after 7th iteration, indicated with vertical line.

Figure 3 indicates that skill level for Djokovic converges to stationary distribution after 7 **iterations**.

Auto-correlation Figures 4 and 5 look at Pearson Correlation coefficient between samples and their preceding samples in a sequence. Figure 5 shows that, for Djokovic, samples appear to be dependent on only 7 **preceding samples**. While this correlation is computed for the entire sequence, and the dependence on a poor initialisation may stay present for more than 7 iterations, Auto-correlation does support argument of **high mixing-rate** and **short Burn-in**.

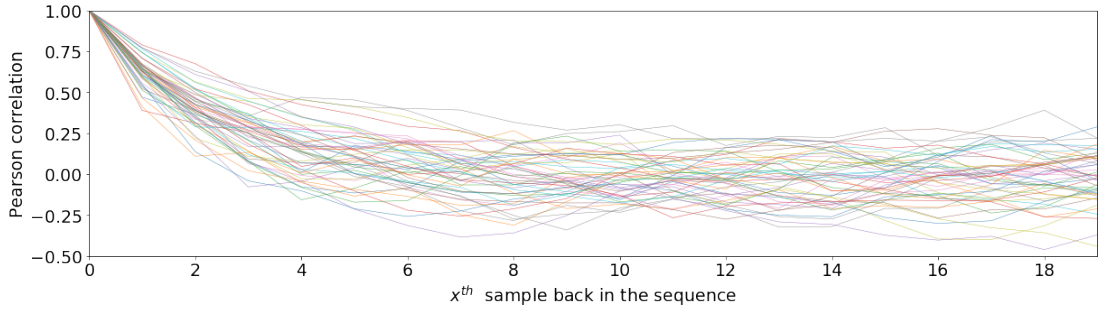


Figure 4: Auto-correlation plot for 50 trials of MC simulation for Novak Djokovic skills.

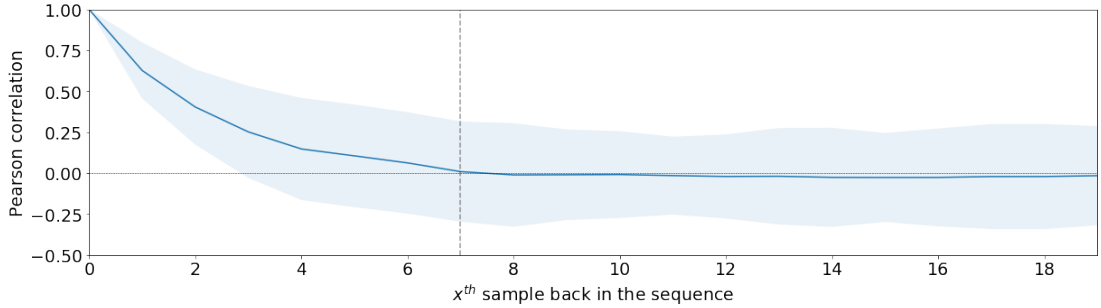


Figure 5: Mean and 95% confidence bar for 50 trials of MC for Auto-correlations from Figure 4. Convergence seen after 7th iteration, indicated with vertical line.

1.4 Conclusion

Figure 6 shows that 96% of players' skills have shorter chain dependence than 7, supporting initial hypothesis that studying convergence of best players can be a proxy for convergence of Gibbs sampling. However, to be conservative, we use Burn-in of $3 \times 7 = \mathbf{21}$.

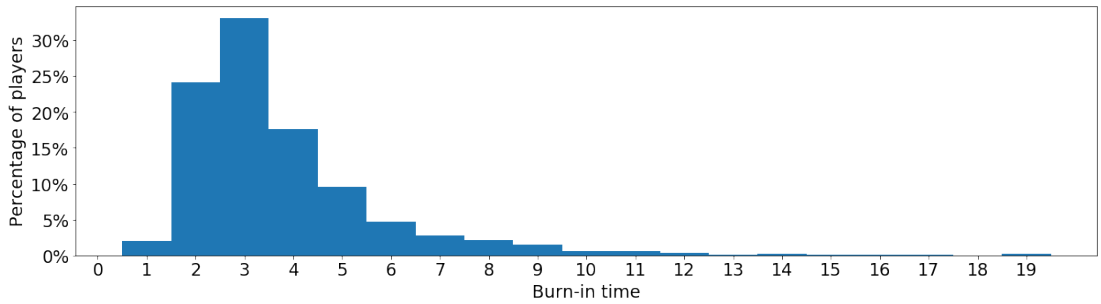


Figure 6: Histogram of Auto-correlation times for 5070 players (107 players in 50 Monte Carlo trials). Figure shows that Auto-Correlation time equal to 7 is conservative.

2 Part B

We want to sample from a joint posterior distribution of player skills given observed data $p(\boldsymbol{w}, t|y)$. However the likelihood term of that distribution doesn't have closed form, which makes the joint intractable, motivating use of approximate methods.

Conditional distributions in **Gibbs sampling** will converge to the joint posterior distribution if it is able to move to a stationary distribution which corresponds to a region of high density of the joint distribution of skills. In Figures 6 and 3 we saw evidence for high mixing rates and convergence to a stationary distribution after 7 iterations.

In **Message passing** structure of the factor graph induces conditional independence of skill marginal distributions for each player given game outputs $p(w_i|y)$. Because of multiple games for some players, the factor graph forms loops and multiple iterations are required to achieve convergence. Figure 7 shows that after 10 iterations marginal distribution of Djokovic's skill comes within 5% of the asymptote of stationary distribution. Message Passing converges to conditionally independent marginal distributions of each player's skill. Players' skill are correlated, however they are conditionally independent given game outcomes.

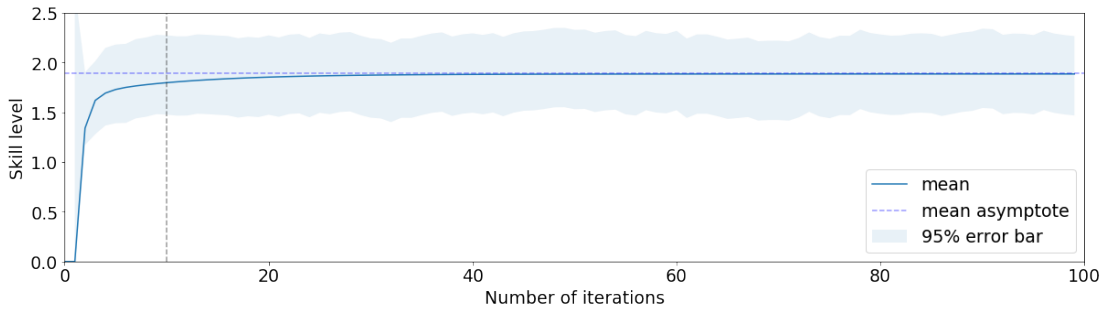


Figure 7: Convergence of conditionally independent marginal distribution of Novak Djokovic's skill, using Expectation Propagation (Message passing) algorithm.

3 Part C

3.1 Table 1

Equation 4 shows derivation of mean and variance of difference of skills s. Since players skills are drawn from marginals, term $Cov[w_1, w_2] = 0$. Equation 7 shows corresponding PDF and CDF used to compute Table 1.

$$\begin{aligned} s &= w_1 - w_2 \\ E[w_1 - w_2] &= E[w_1] - E[w_2] \\ Var[w_1 - w_2] &= Var[w_1] + Var[w_2] - 2Cov[w_1, w_2] \end{aligned} \tag{4}$$

$$\begin{aligned} p(s|w_1, w_2) &= N(s; w_1 - w_2, \sigma_1^2 + \sigma_2^2) \\ p(s > 0|w_1, w_2) &= 1 - CDF(0, \mu = w_1 - w_2, \sigma = \sqrt{\sigma_1^2 + \sigma_2^2}) \end{aligned} \tag{5}$$

	Nadal	Federer	Murray	Djokovic
Nadal	-	0.43	0.77	0.06
Federer	0.57	-	0.81	0.09
Murray	0.23	0.19	-	0.01
Djokovic	0.94	0.91	0.99	-

Table 1: Probability that the skill of player in the first column, is higher than the skill of players' in subsequent columns.

3.2 Table 2

For the probability of game outcome, in addition to s, we also account for the performance noise $n \sim p(n; 0, 1)$. Equation 6 shows how noise n increases the variance.

$$\begin{aligned} t &= w_1 - w_2 + n \\ E[w_1 - w_2 + n] &= E[w_1] - E[w_2] \\ Var[w_1 - w_2] &= Var[w_1] + Var[w_2] + 1 \end{aligned} \tag{6}$$

$$\begin{aligned} p(t|w_1, w_2) &= N(t; w_1 - w_2, 1 + \sigma_1^2 + \sigma_2^2) \\ p(y|w_1, w_2) = p(t > 0|w_1, w_2) &= 1 - CDF(0, \mu = w_1 - w_2, \sigma = \sqrt{(1 + \sigma_1^2 + \sigma_2^2)}) \end{aligned} \tag{7}$$

	Nadal	Federer	Murray	Djokovic
Nadal	-	0.48	0.57	0.34
Federer	0.52	-	0.59	0.36
Murray	0.43	0.41	-	0.28
Djokovic	0.66	0.64	0.72	-

Table 2: Probability that player in the first column will win a game with players in other columns.

3.3 The difference

The extra performance noise term increases variance of $p(t > 0|w_1, w_2)$, making the game outcome more uncertain. Because tails of a gaussian have positive curvature, the mass density shifts away from the mean, making CDF more favourable for the worse player, than the CDF of difference of skills is. It accounts for worse players being able to get "lucky" when some external factors, which are not directly related to players skills, affect the game outcome in their favour.

4 Part D

4.1 Method 1

Figure 8 shows Gaussian fit to 427 thinned skill samples for each player.

$$p(s > 0|w_D, w_N) = 1 - CDF(0, \mu = w_D - w_N, \sigma = \sqrt{\sigma_D^2 + \sigma_N^2}) = 0.926 \tag{8}$$

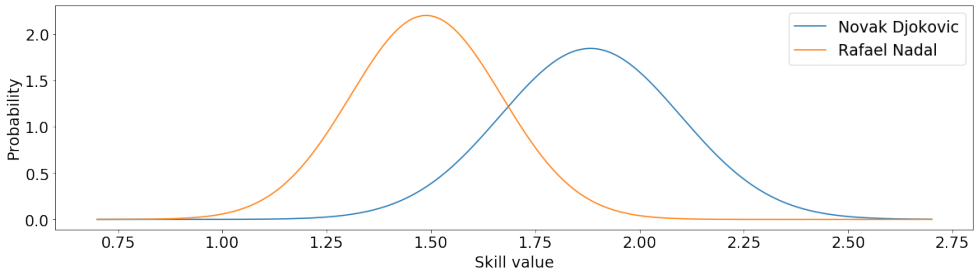


Figure 8: Approximated Marginal distributions over skills for Djokovic and Nadal. Constructed using 427 Gibbs samples, sampled in intervals of 7 for sample independence.

4.2 Method 2

$p(s > 0|w_D, w_N)$ estimate in Method 1, can be improved by accounting for non-zero covariance term of joint distribution, which Gibbs sampling approximates.

$$\mu = \begin{bmatrix} \mu_D \\ \mu_N \end{bmatrix} = \begin{bmatrix} 1.90 \\ 1.48 \end{bmatrix} \tag{9}$$

$$\mathbf{\Sigma} = \begin{bmatrix} \Sigma_{DD} & \Sigma_{DN} \\ \Sigma_{ND} & \Sigma_{NN} \end{bmatrix} = \begin{bmatrix} 0.045 & 0.014 \\ 0.014 & 0.039 \end{bmatrix} \tag{10}$$

$$\begin{aligned} p(s|w_D, w_N) &= N(s; w_D - w_N, \Sigma_N + \Sigma_D - 2\Sigma_{ND}) \\ p(s > 0|w_D, w_N) &= 1 - CDF(0, \mu = w_D - w_N, \sigma = \sqrt{\Sigma_D + \Sigma_N - 2\Sigma_{ND}}) = 0.961 \end{aligned} \tag{11}$$

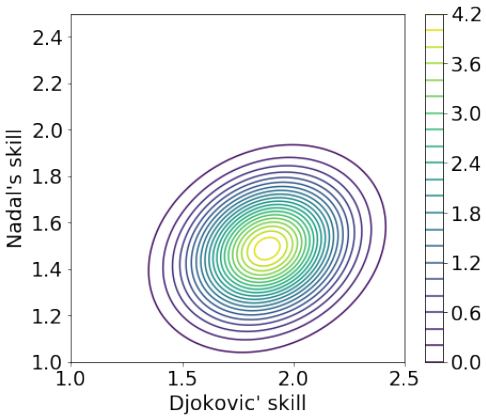


Figure 9: Join distribution between Nadal's and Djokovic's skill computed based on Gibbs samples as defined in Figure 8

4.3 Method 3

In Table 3, we have directly counted number of positive skill differences $w_1 - w_2$ for all sample pairs. By keeping samples in pairs, we can utilise information about their covariances.

	Djokovic	Nadal
Total won	408	18
Fraction won	0.958	0.042

Table 3: Out of 427 skill sample pairs considered, Djokovic's skill was higher 408 times.

4.4 Best method

Method 1 disregards information about covariance of Players' skills. Every time two players play together, their joint distribution variances shrink and covariances grow. We become more certain about their individual future games, but even more so, about the future games they play together. Methods 2 and 3 both account for this information.

We expect **Method 3** to provide more accurate probability than Method 2, because it doesn't approximate non-Gaussian joint distribution with a Gaussian (see Part B). However, the information loss due to approximation is small, $(0.961 - 0.958 = 0.003)$, and **Method 2** allows for simple representation of uncertainty in form of covariance matrix, necessary for Bayesian inference, which makes **Method 2** preferable.

4.5 Table - Method 2

Both Tables 4 and 1 (Part C), lead to the same rankings of top 4 players. Probabilities in these tables are on average only 0.010 different, suggesting that both methods approximate same underlying distribution.

	Nadal	Federer	Murray	Djokovic
Nadal	-	0.42	0.8	0.04
Federer	0.58	-	0.82	0.08
Murray	0.2	0.18	-	0.01
Djokovic	0.96	0.92	0.99	-

Table 4: Computed using CDF in **Method 2**. Table shows probability that skill of player in the first column will be higher than skill of players in other columns.

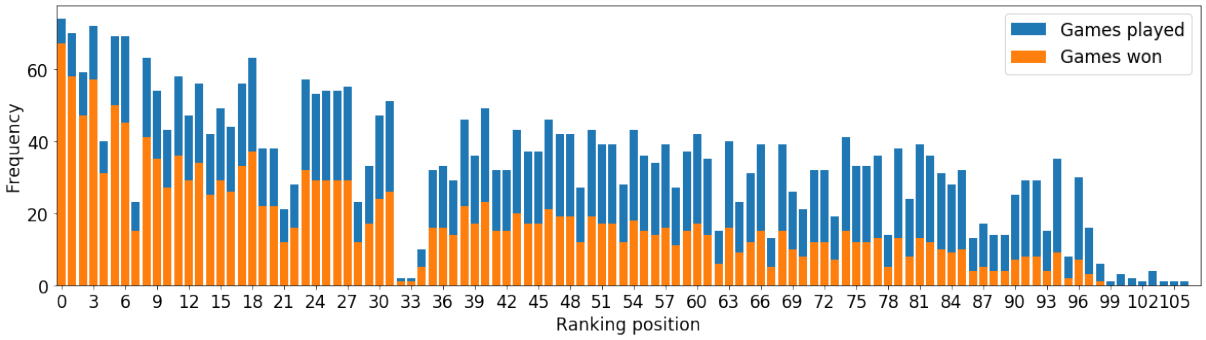
5 Part E

For plots of each ranking see Appendix.

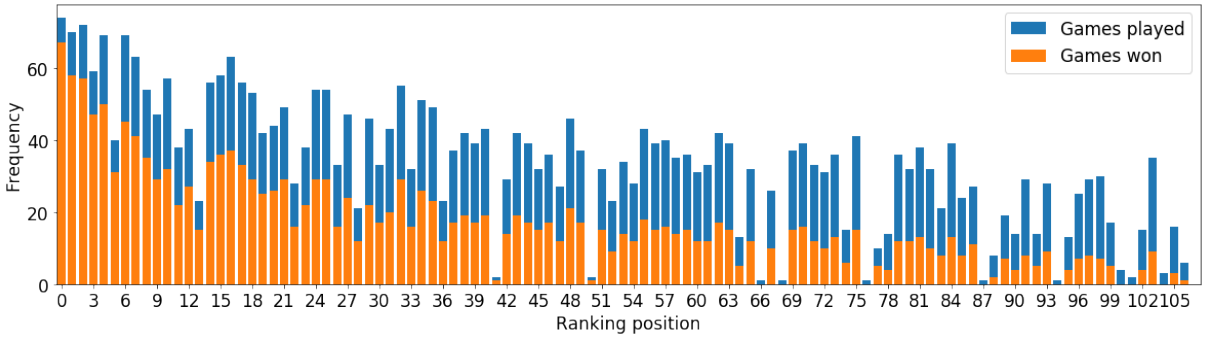
5.1 Frequentist vs. Bayesian

Frequentist approach of Win-rate ranking doesn't account for skill uncertainty of players, especially those with little game history. Figure 10.a shows that players with no won games fill the end of the ranking, as their win-ratio equals 0. An alternative failure of frquentist approach would be to rank Top 1, a player who won the only game he played.

Both Gibbs and MP rankings take **Bayesian** approach to inference. They model high skill variance for players with little game history, to account for our uncertainty about their skill.



(a) Win-rate ranking



(b) Gibbs ranking

Figure 10: Count of played and won games for players sorted using rankings specified in sub-captions.

5.2 Gibbs vs. MP

Gibbs sampling and MP rank 41 players at exactly same position, with another 41 being shifted only by a single rank, Figure 11. Both rankings are similar, what can be also observed from modeled players' skill means, Figure 12 and variances.

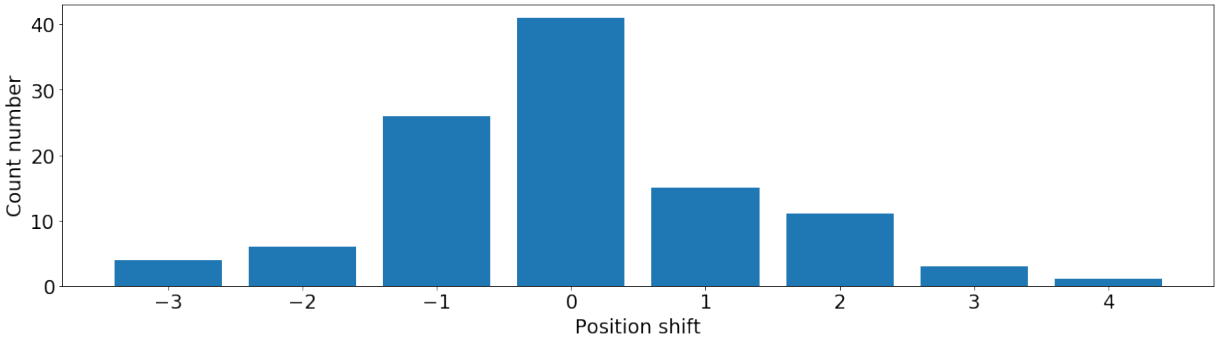


Figure 11: Shift in position of players in MP ranking as compared to Gibbs ranking.

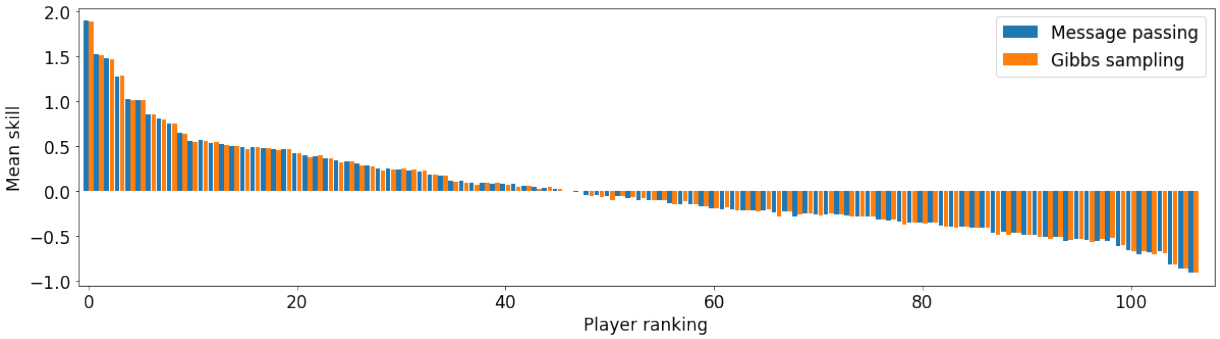


Figure 12: Comparison of players' skill means for Gibbs and MP approximations of posterior distribution, sorted using Gibbs ranking.

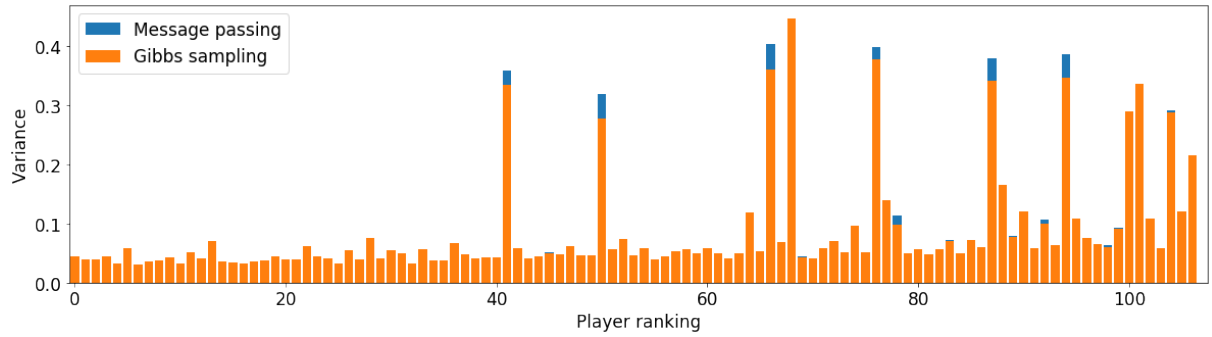
Looking at Figure 13, we see that Gibbs ranking predicts higher variance for top players, while Message passing for those who played few games (high variance bars).

If we look at Figure 14 we see that top players, in Gibbs model, have high ratio of positive to negative covariances, which will decrease the performance variance in Equation 12. Thus, decreasing effect of the gap between blue and orange bars in 13.b.

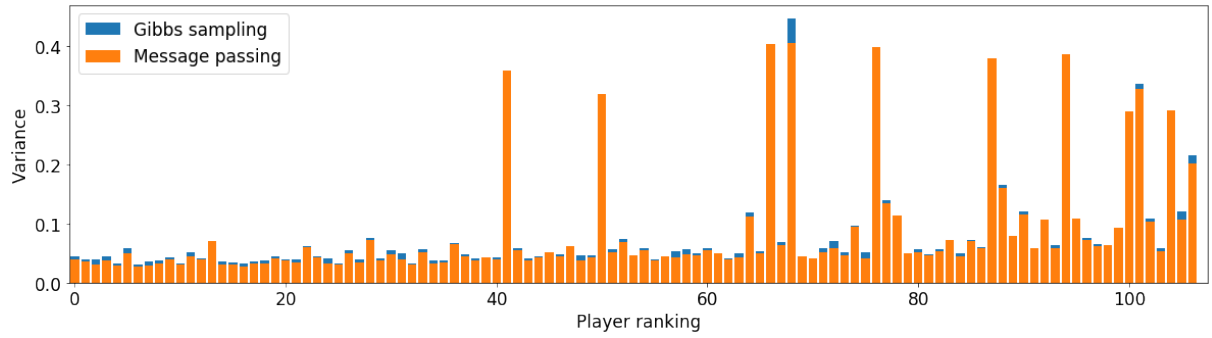
Similarly, lower-ranked players in Figure 14, have high ratio of negative to positive covariances, which will increase Σ_t and decrease the effect of variance gap in Figure 13.a.

$$\Sigma_t = \Sigma_{w_i} + \Sigma_{w_j} + 1 - 2 \times Cov(w_i, w_j) \tag{12}$$

This shows that both methods have converged to similar approximations of posterior distribution. In practice, we may prefer MP due to computational efficiency.



(a) Message passing variance underneath



(b) Gibbs sampling variance underneath

Figure 13: Overlayed variances of players' skill for both approximation methods and sorted using Gibbs ranking.

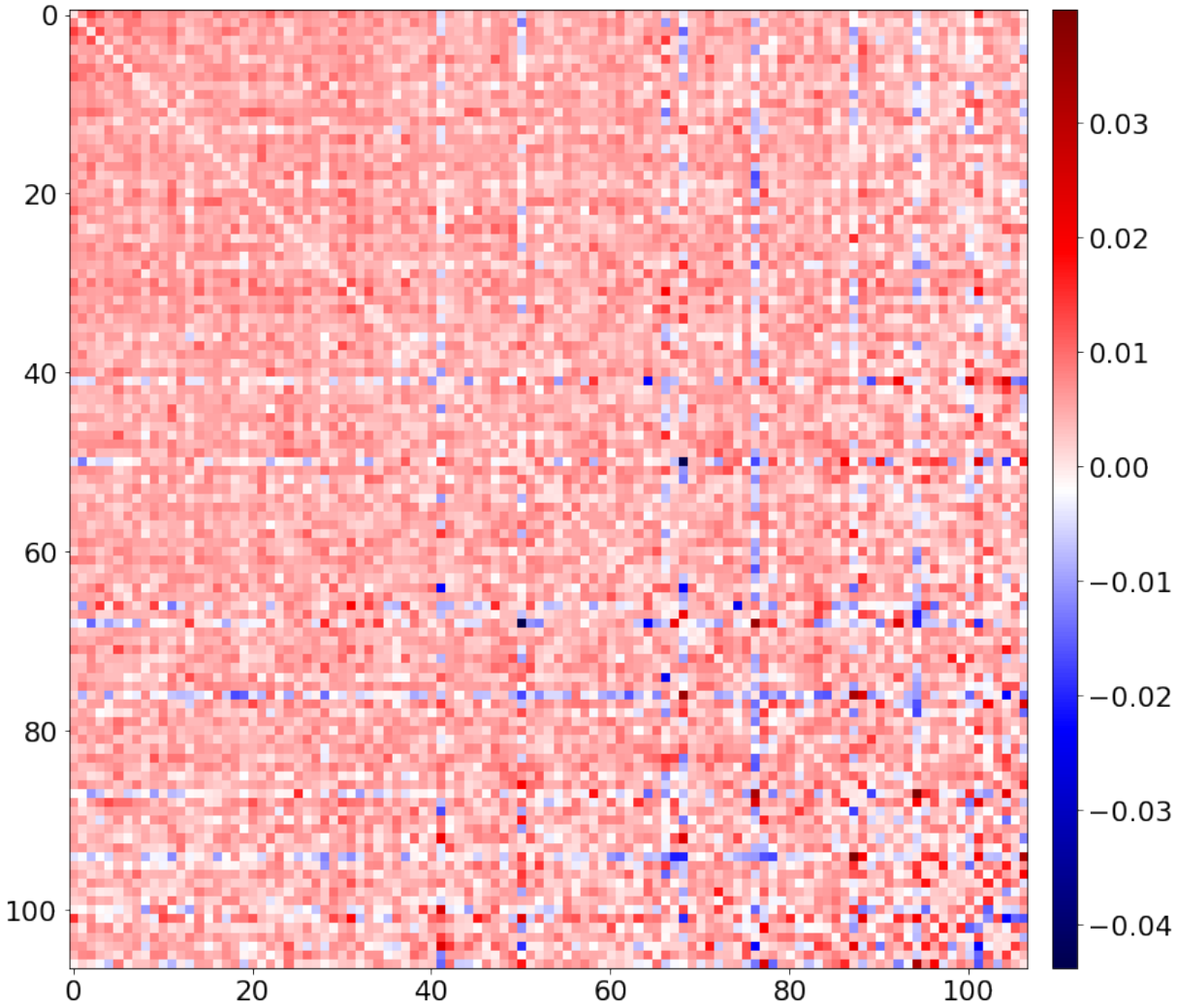


Figure 14: Covariance matrix of players sorted according to Gibbs ranking. Diagonal variances are masked with 0 for readability of the color bar. Red color coresponds to positive values, while blue to negative values.

Appendix A

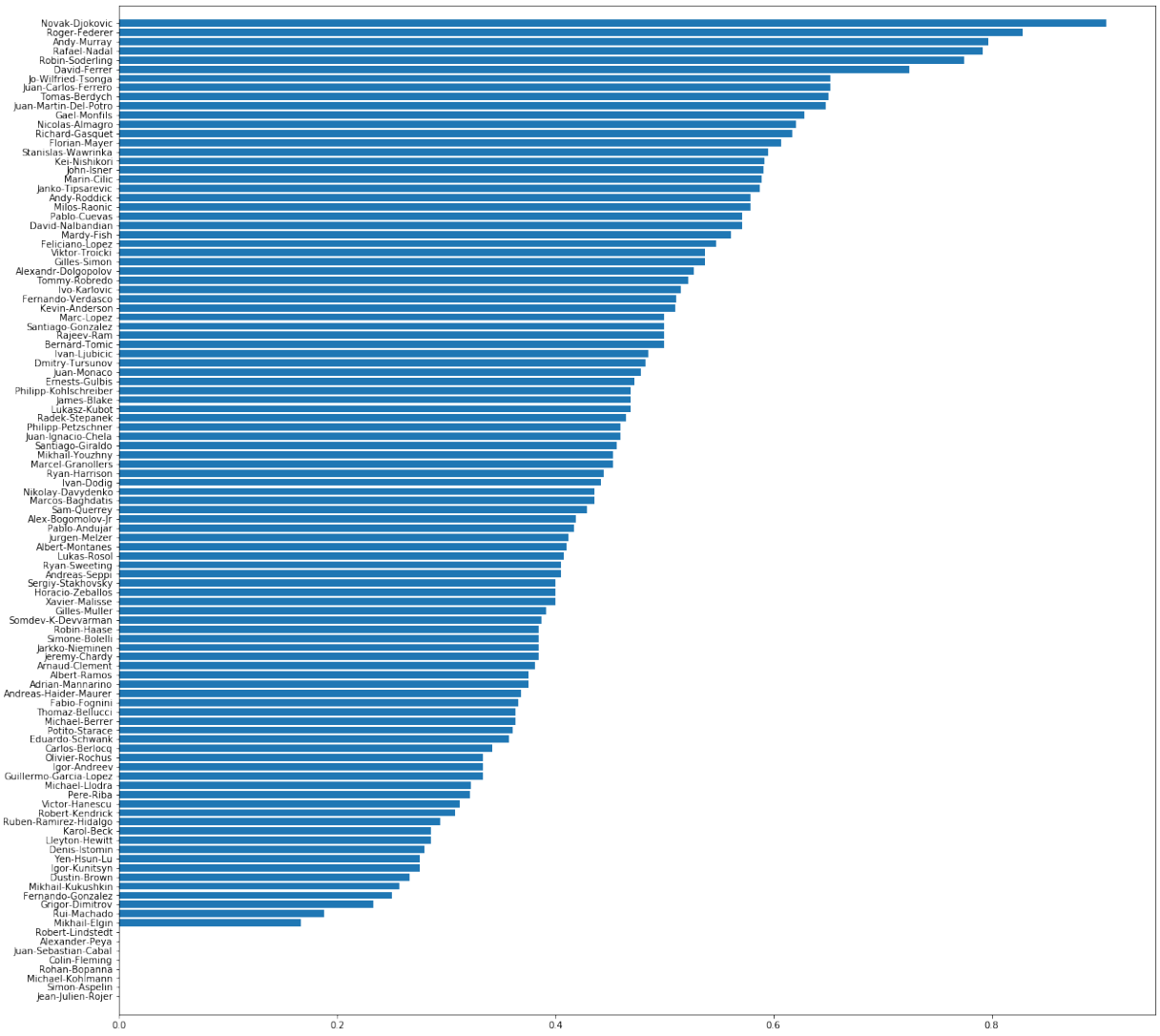


Figure 15: Ranking computed by comparing player win rates based on observed game outcomes.

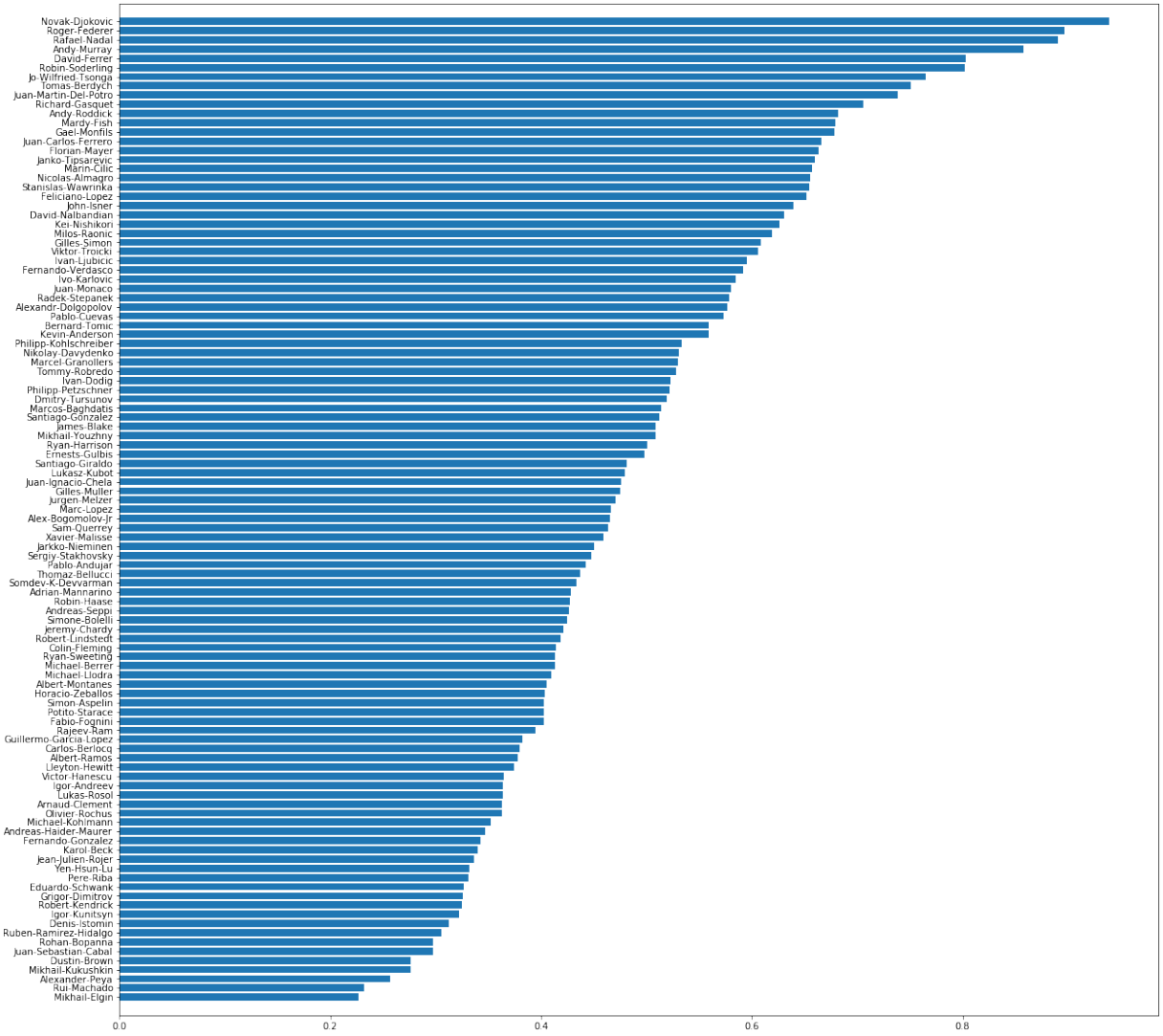


Figure 16: Ranking made by averaging predicted player performance with all other players in the ranking, where skill levels were modeled using joint distribution of Gibbs samples.

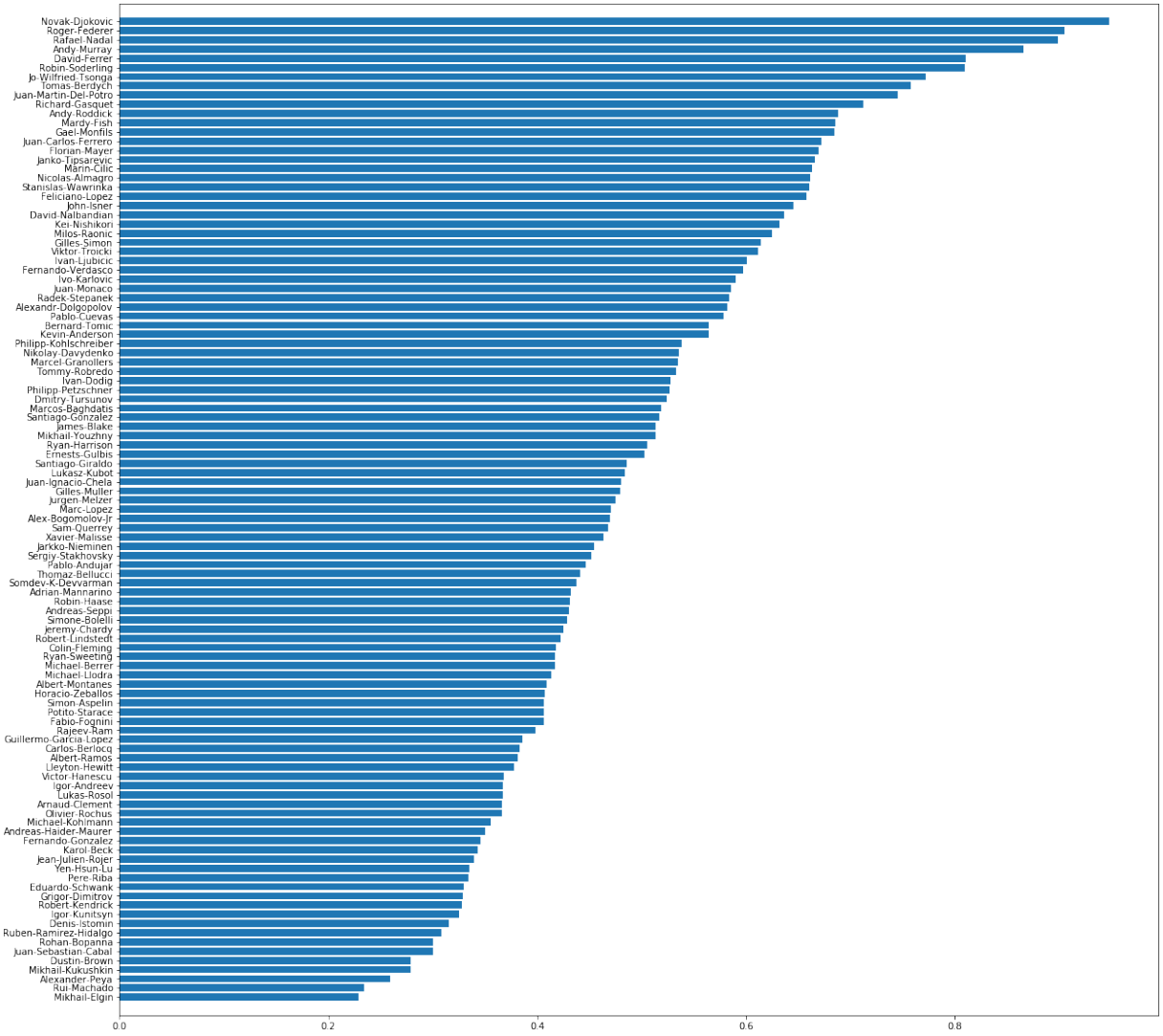


Figure 17: Ranking made by averaging predicted player performance with all other players in the ranking, with skills modeled by conditionally independent marginal distributions computed using Message passing algorithm.