4F13 Probabilistic Machine Learning
Coursework #2: Probabilistic Ranking

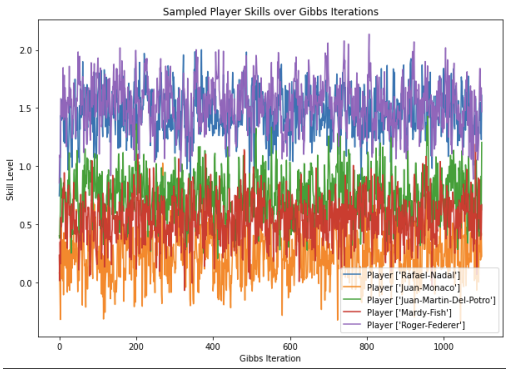
Candidate Number: 5504C
Word count: 972

November 17, 2023

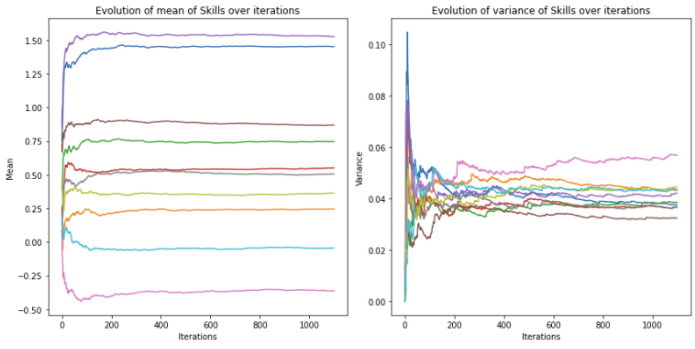
1 Exercise a)

The Gibbs sampling process involves iterative sampling from the conditional distributions of player skills (Figure.1a), where convergence to the target distribution is indicated by stable mean and variance of the skill estimates (Figure.1b). The burn-in period is empirically determined from these plots, where the transient effects of initial conditions dissipate, and the samples begin to reflect the true underlying distribution. This is typically the flat region on the mean plot, which, based on the provided data, appears to be reached after approximately 190 iterations. Autocorrelation times are assessed by examining the decay of the autocorrelation function (ACF) with increased lag (Figure.2). The ACF should ideally fall within the confidence bounds near zero for samples to be considered independent. From Figure.2, we observe that the coefficients approach zero after a lag of approximately 5, suggesting that subsequent samples beyond this lag are nearly independent.

Given these observations, a conservative estimate for the minimum number of Gibbs iterations required for reliable skill estimates would be the lag at which autocorrelation becomes negligible, multiplied by a factor to account for statistical certainty, resulting in over 1000 iterations. This ensures that the autocovariance of the skill samples is minimized, allowing for independent and identically distributed samples from the posterior distribution.



(a) Gibbs sample of skills for some of the players



(b) Convergence of mean and variance of the sampled skills over iterations

Figure 1: Gibbs sampling skills and their mean and variance convergence for 10 players

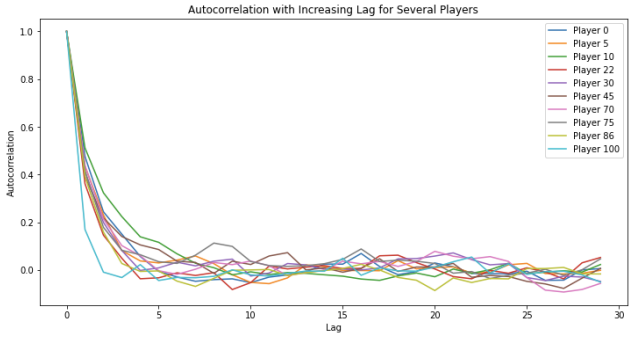


Figure 2: Autocorrelation function for 10 random players with increasing lag

2 Exercise b)

Expectation Propagation (EP) refines marginal skill distributions through iterative message passing. Convergence is achieved when the updated distributions align with the empirical data, as shown in Figure.3. The figure shows different convergence of means and variances for 10 players, and each player's skill distribution converges at a distinct rate. The precisions stabilize quickly, generally within 4 iterations. In contrast, the means tend to require up to 45 iterations before it reaches a comparable state of stability.

In Gibbs sampling, convergence is achieved when the Markov chain converges to the stationary distribution, which is the true underlying joint distribution we aim to characterize. This convergence indicates that the Markov chain has forgotten its initial state and the samples are now representative of the entire distribution space. Message passing algorithms like Expectation Propagation (EP) converge when the iterative exchange of messages between nodes in a factor graph leads to stable marginal distributions. Unlike the Gibbs sampler, which directly samples from the joint distribution, message passing works by refining beliefs about each variable's distribution through local updates.

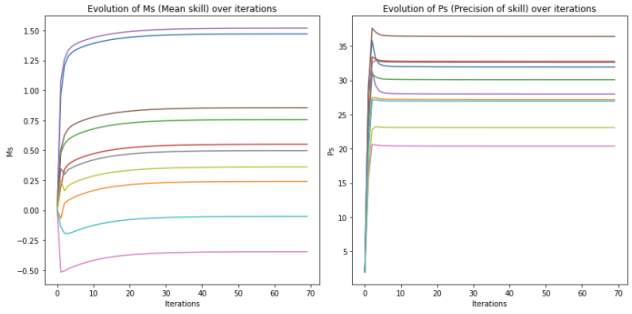


Figure 3: Convergence of mean and variance of skills using message passing algorithm

3 Exercise c)

The first table represents the probabilities computed from EP that the skill level of one player is higher than another for the top 4 players according to ATP ranking. For instance, the probability that Novak Djokovic has a higher skill level than Rafael Nadal is 0.940. This is calculated by treating the difference in skill levels as a random variable and computing the probability that this difference is greater than zero with $\sim \mathcal{N}\left(\mu_1 - \mu_2, \frac{1}{\text{precision}_1} + \frac{1}{\text{precision}_2}\right)$:

$$P(S = \text{skill}_1 - \text{skill}_2 > 0) = 1 - \Phi(z_s) = 1 - \Phi(\frac{s - \mu_s}{\sigma_s}) = 1 - \Phi(\frac{0 - \mu_s}{\sigma_s})$$

The second table shifts the focus to the probability of one player winning a match against the other, rather than just having a higher skill level. The probabilities here take into account not only skill levels but also the performance variability that can influence match outcomes, and the variance of the distribution becomes ($\frac{1}{\text{precision}_1} + \frac{1}{\text{precision}_2}$ + performance Variance). This encompasses performance variance in players. Therefore, while Djokovic's skill is rated significantly higher than Nadal's, the probability of Djokovic winning a match is a more modest 0.655, indicating that the performance factor plays a crucial role in the outcome of a match.

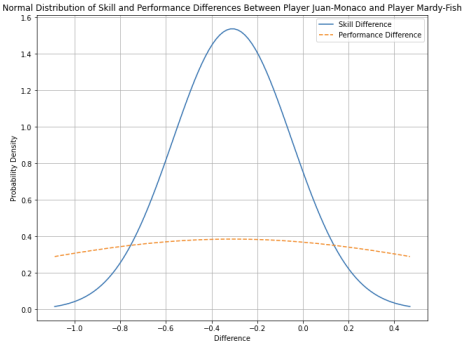
The difference between the two tables lies in the parameters they evaluate: the first table is concerned solely with **skill differences as a static attribute**, while the second table considers the **dynamic and unpredictable nature of match performance**, leading to the more spread distribution and probabilities indicated by Figure.4.

	B			
A	Novak-Djokovic	Rafael-Nadal	Roger-Federer	Andy-Murray
Novak-Djokovic	-	0.060	0.091	0.015
Rafael-Nadal	0.940	-	0.573	0.234
Roger-Federer	0.909	0.427	-	0.189
Andy-Murray	0.985	0.766	0.811	-

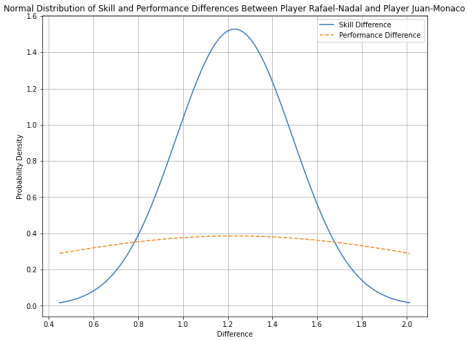
Table 1: Probability that player B has a higher skill than player A ($p_B > p_A$).

	B			
A	Novak-Djokovic	Rafael-Nadal	Roger-Federer	Andy-Murray
Novak-Djokovic	-	0.345	0.362	0.280
Rafael-Nadal	0.655	-	0.518	0.427
Roger-Federer	0.638	0.482	-	0.409
Andy-Murray	0.720	0.573	0.591	-

Table 2: Probability that player B has higher performance than player A ($p_B > p_A$).



(a) Skill difference and performance difference for Juan-Monaco and Mardy-Fish



(b) Skill difference and performance difference for Rafael-Nadal and Juan-Monaco

Figure 4: Skill difference and performance difference distribution of two of the games

4 Exercise d)

When evaluating the skills of Nadal and Djokovic using Gibbs sampling, several methods can be employed. The first method involves utilizing the skill marginals, which are derived by taking the mean and variance of the skill samples and assuming a normal distribution for each player's skills. This method is plotted in Figure.5a showing the marginal Gaussian distributions, where the probability of Djokovic having a higher skill than Nadal is calculated to be 0.936.

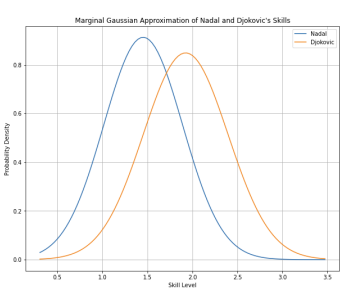
The second method considers the joint skill distribution, which incorporates both the individual skill levels and the covariance between players, thus providing a more holistic view of the skill relationship. The contour plot of the joint Gaussian distribution illustrates how the covariance affects the probability calculations, with the probability of Djokovic's skill being greater than Nadal's slightly higher at 0.963, considering the covariance between their skills.

A direct comparison using samples can offer a more immediate approach, as shown in the histogram plot, which represents the skill levels distribution from collected samples. This direct method yields a probability of 0.948 for Djokovic's higher skill, suggesting that, when sufficient samples are considered, the direct method can be quite definitive.

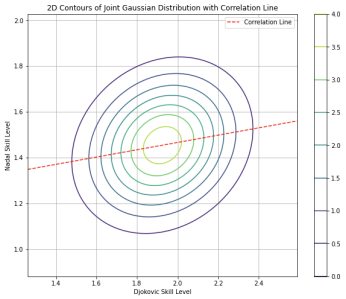
While using marginal distributions or direct samples is computationally faster, it may neglect the additional information provided by the covariance in the joint distribution method. The joint Gaussian approach, which **accounts for covariance**, offers a more nuanced probability and can smooth out the variability that may arise from a direct sample comparison, reducing the noise inherent in the sampling process. And the computed probabilities of skill difference is shown in Table.4.

Property	Method 1 (Marginal)	Method 2 (Joint)	Method 3 (Direct)
Mean (Nadal)	1.454	-	-
Variance (Nadal)	0.191	-	-
Mean (Djokovic)	1.925	-	-
Variance (Djokovic)	0.220	-	-
Mean (Joint)	-	$[1.925, 1.454]$	-
Covariance Matrix	-	$\begin{bmatrix} 0.049 & 0.007 \\ 0.007 & 0.037 \end{bmatrix}$	-
$P(s_D > s_N)$	0.936	0.963	0.948

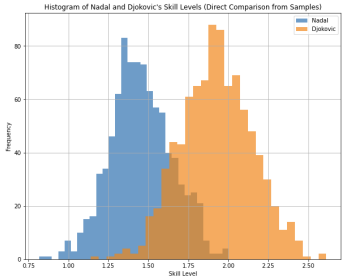
Table 3: Properties of the three methods



(a) Marginal Gaussian distributions for individual player skills



(b) Joint Gaussian distribution considering skill covariance



(c) Histogram of skill levels from direct sample comparisons

Figure 5: Comparative analysis of different methods for skill level estimation.

	B			
A	Novak-Djokovic	Rafael-Nadal	Roger-Federer	Andy-Murray
Novak-Djokovic	-	0.037	0.062	0.013
Rafael-Nadal	0.963	-	0.614	0.248
Roger-Federer	0.938	0.386	-	0.197
Andy-Murray	0.987	0.752	0.803	-

Table 4: Probability that player B has a higher skill level than player A ($p_B > p_A$).

5 Exercise e)

The empirical method involves calculating the average outcomes of games directly from historical data. This approach takes a straightforward average of the outcomes, providing a ranking based on observed wins and losses shown in Figure.6. The empirical ranking method is susceptible to variance due to small sample sizes or atypical results, which can skew the averages, particularly for players with fewer games.

Gibbs sampling excels in handling complex, multi-dimensional probability distributions and incorporating latent variables like skill level differences. The ranking is shown in Figure.7. However, the method is computationally intensive and converges slowly in high-dimensional spaces. Additionally, the accuracy of Gibbs sampling-based predictions heavily depends on the chosen model and its assumptions, which may not always accurately capture the real-world games.

The message passing algorithm updates the marginal skill distributions iteratively. The ranking is shown in Figure.8 This method incorporates understandings of the network of player matchups and can refine predictions as more data is propagated through the system. It provides exact results in tree-structured graphs, but in loopy graphs, they rely on approximations, which might lead to inaccuracies.

In conclusion, empirical game outcome averages can bias rankings towards players with fewer games by overestimating their winning probabilities, whereas probabilistic methods like Gibbs sampling and message passing algorithms offer more nuanced predictions by considering a broader data set and the inter-connections of player performances.

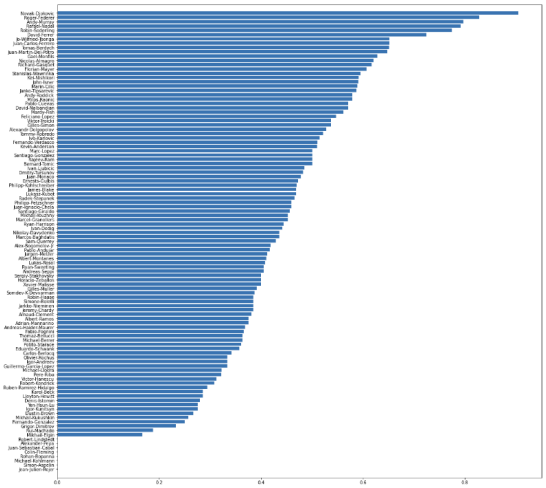


Figure 6: Ranking based on empirical method

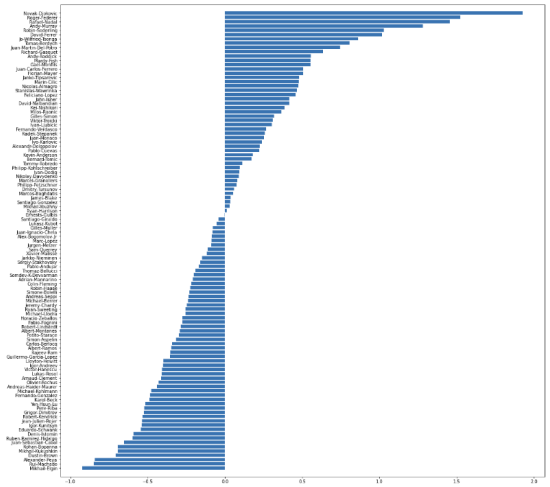


Figure 7: Ranking based on Gibbs sampling

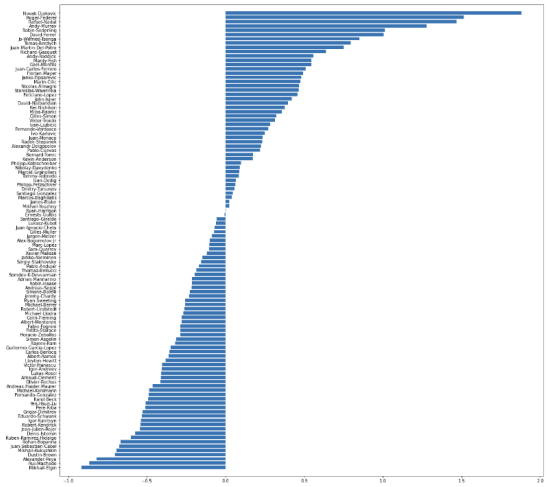


Figure 8: Ranking based on message passing

6 Appendix

6.1 (a)

```
for p in range(M):
    m[p] =  sum(t[np.where(G[:, 0] == p)]) - sum(t[np.where(G[:, 1] == p)])
...
for g in range(N):
    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
```

6.2 (c)

```
def higher_skill(player1, player2, means, precisions, perf_var=0):
    mean = means[player1, -1] - means[player2, -1]
```

```
    var = 1/precisions[player1, -1] + 1/precisions[player2, -1] + perf_var
    prob = 1- norm.cdf((0-mean)/(var**0.5)) # Z-score normalization to turn the horizontal
    axis to sd
    return prob
```

6.3 (e)

```
sorted_barplot(rank_empirical, W)

means_gibbs_skills = np.mean(skill_samples[:,burn_in:],axis=1)
sorted_barplot(means_gibbs_skills,W)

sorted_barplot(mean_player_skills[:, -1], W)
```