

# 4F13 Coursework 2 - Probabilistic Ranking

November 2024

## 1 Task A

Figure 1 shows samples of the skills at each Gibbs iteration for 3 players. We see that at each iteration the sampled skill of each player appears random but perhaps not independent of the previous iteration. This is expected from an MCMC method like Gibbs sampling and shown better by the auto covariance plot in Figure 2.

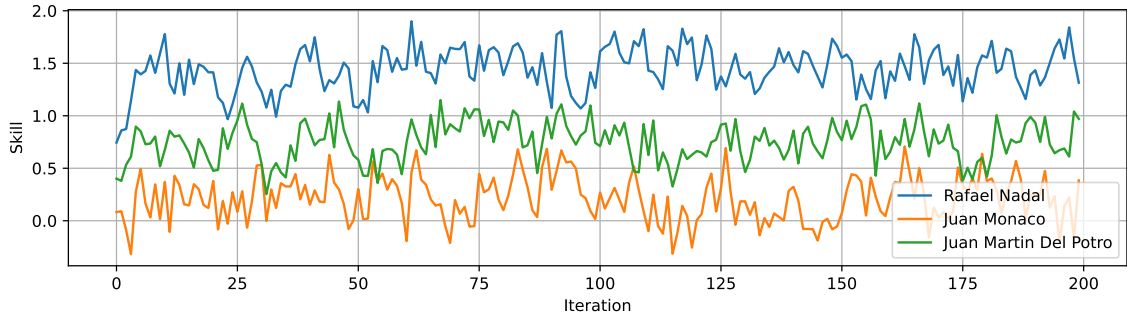


Figure 1: Samples of the skill of 3 players at each Gibbs iteration.

If the samples at each iteration were independent we would expect the auto covariance to be zero for all non zero lags. However, we notice that the auto covariance only converges to zero for all players for lags greater than 10. This indicates that if we kept only every 10th sample they would be independent, this is the logic behind thinning. We do not need to employ thinning techniques here as we are not running into memory issues but it is a useful tool for large datasets.

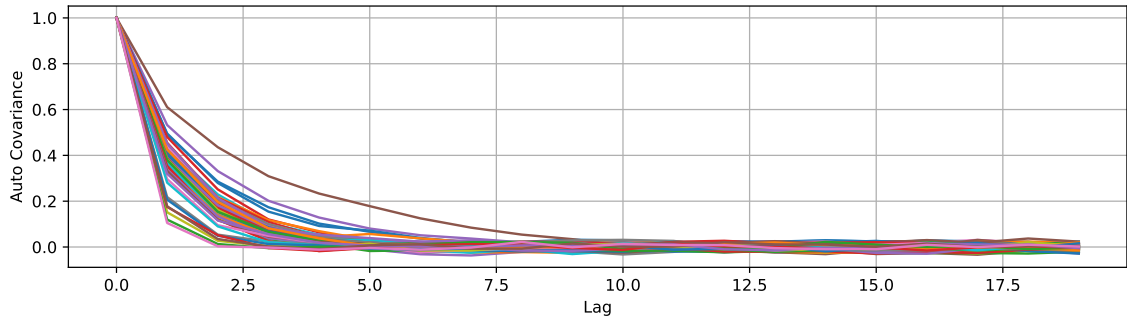


Figure 2: Skill sample auto covariance for all players

The convergence of the skill population is shown in Figure 3. We see that there is no clear burn in time for the population mean but the standard deviation takes about 5 iterations to enter the high probability region. This makes sense from our auto covariance analysis as we concluded samples 10 iterations apart are roughly independent so the burn in time should be no more than 10 iterations. For any further we will discard the first 100 Gibbs iterations as burn in time to be sure.

To estimate how many iterations we need to run the Gibbs sampler to get skill estimates within a certain tolerance we can analyse the variance of our Monte Carlo skill estimates. Equation 1 calculates the variance of the Gibbs skill estimate,  $\bar{w}_i$ , for player  $i$  after  $N$  iterations. The variance depends on the sum of the auto covariance,  $\frac{\sigma_i^{(t)2}}{\text{Var}(w_i)}$  and is inversely

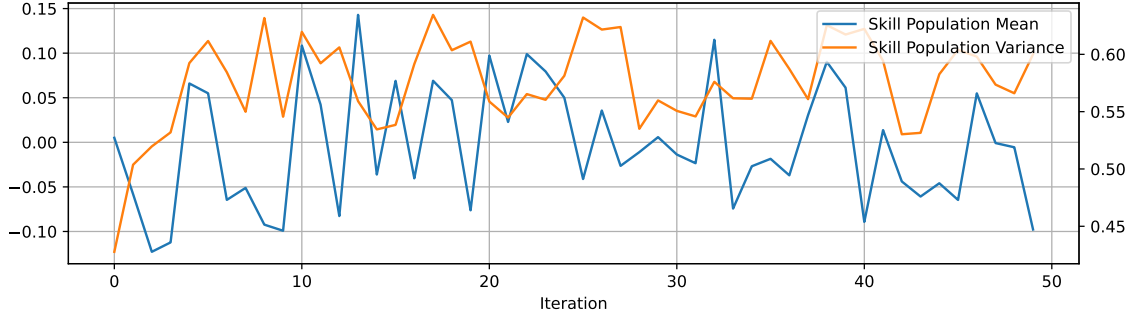


Figure 3: Skill populations mean and standard deviation with Gibbs iteration.

proportional to  $N$ . We compute the average of this variance increase term over all players from the data in Figure 2 and find a variance increase factor of 4. From these calculation we find that in order to achieve a skill estimate variance  $10^{-3}$  smaller than our posterior skill variance we must run our Gibbs sampler for 4000 iterations.

$$\begin{aligned} \text{Var}(\bar{w}_i) &= \frac{1}{N} \text{Var}\left(\sum_{n=1}^N w_i^{(n)}\right) = \frac{\text{Var}(w_i)}{N} \left(1 + 2 \sum_{t=1}^{N-1} \left(1 - \frac{t}{N}\right) \frac{\sigma_i^{(t)2}}{\text{Var}(w_i)}\right) \\ &\approx \frac{\text{Var}(w_i)}{N} \left(1 + 2 \sum_{t=1}^{\infty} \frac{\sigma_i^{(t)2}}{\text{Var}(w_i)}\right) \approx 4 \frac{\text{Var}(w_i)}{N} \end{aligned} \quad (1)$$

## 2 Task B

In the Gibbs sampling approach we are sampling from a Markov chain with stationary distribution  $p(\mathbf{w}|\mathbf{y})$  and we wish to compute expectations of the form  $\mathbb{E}_{\mathbf{w} \sim p(\mathbf{w}|\mathbf{y})}[f(\mathbf{w})]$ . The distribution of our samples must first converge to the stationary distribution, the burn-in period, and then our expectation estimate needs to converge too. From the previous section we know that the variance of the expectation is proportional to  $N^{-1}$ .

With message passing we are converging to a stable graph with the distribution of variables at the nodes being approximated by Gaussians. We can say that the graph has converged within a tolerance  $\epsilon$  at iteration  $n$  when the absolute change of all marginal parameters between iterations are less than  $\epsilon$ . Equation 2 formalises this with  $m_i^{(n)}$  and  $p_i^{(n)}$  being the estimated mean and precision of player  $i$  at iteration  $n$ .

$$|m_j^{(n)} - m_j^{(n-1)}| < \epsilon \quad \text{and} \quad \max_i |p_i^{(n)} - p_i^{(n-1)}| < \epsilon \quad \forall i \quad (2)$$

We plot the largest absolute change between iterations for each parameter in Figure 2. By inspection of Figure 2 we notice that convergence is exponential and we achieve a tolerance of machine precision within 300 iterations. This is much faster than the Gibbs sampler where  $\epsilon \propto \sqrt{\text{Var}(\bar{w}_i)} \propto N^{-0.5}$ .

## 3 Task C

Using the skill parameters estimated by the EP algorithm we denote the skill of player  $i$  as  $w_i \sim \mathcal{N}(m_i, p_i^{-1})$ , with  $m_i$  and  $p_i$  being the mean and precision for player  $i$ . The distributions of the skill difference,  $s_{ij} = w_i - w_j$ , and performance difference,  $t_{ij} = s_{ij} + \eta$  :  $\eta \sim \mathcal{N}(0, 1)$ , between player  $i$  and  $j$  are given in Equation 3. These are simple to compute under the assumption that  $w_i$ ,  $w_j$  and  $\eta$  are independent.

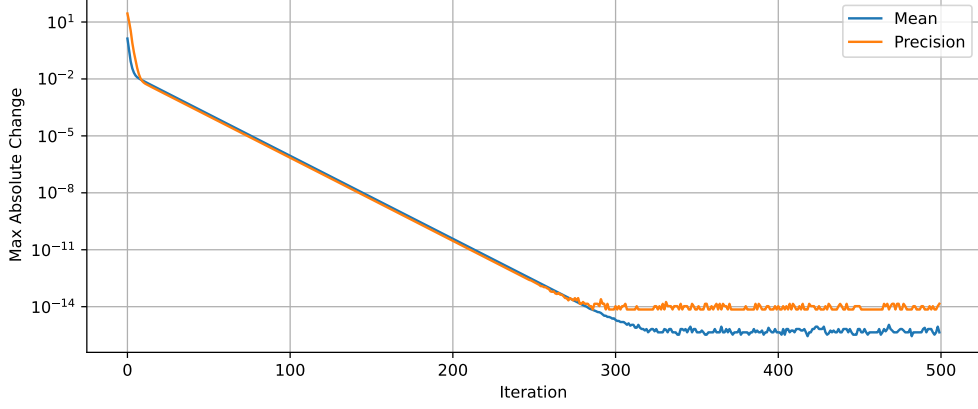


Figure 4: Convergence of the mean and precision of the marginal skill of all players.

$P(s_{ij} > 0)$	Djokovic	Nadal	Federer	Murray
Djokovic	-	0.94	0.91	0.99
Nadal	0.06	-	0.43	0.77
Federer	0.09	0.57	-	0.81
Murray	0.01	0.23	0.19	-

$P(t_{ij} > 0)$	Djokovic	Nadal	Federer	Murray
Djokovic	-	0.66	0.64	0.72
Nadal	0.34	-	0.48	0.57
Federer	0.36	0.52	-	0.59
Murray	0.28	0.43	0.41	-

(a) Probability that row player is more skilled than column player. (b) Probability that row player wins a match against column player.

Table 1: Skill and performance differences between top 4 ATP players based on EP.

$$\begin{aligned}
 s_{ij} &\sim \mathcal{N}(m_i - m_j, p_i^{-1} + p_j^{-1}) \\
 t_{ij} &\sim \mathcal{N}(m_i - m_j, p_i^{-1} + p_j^{-1} + 1)
 \end{aligned} \tag{3}$$

We wish to compute the probability that player  $i$  has greater skill than player  $j$ ,  $P(s_{ij} > 0)$ , and the probability that player  $i$  will win a match against player  $j$ ,  $P(t_{ij} > 0)$ . We use the identity  $P(x > 0) = \Phi(\frac{\mu}{\sigma})$  where  $\Phi$  is the standard normal c.d.f. and  $x \sim \mathcal{N}(\mu, \sigma^2)$ . We find these probabilities between the top 4 ranked ATP players in Table 1.

We are always more confident in who is more skilled than who will win a match. This makes sense as  $\sigma_t > \sigma_s$ , however, intuitively this is also clear as we are always less certain about the outcome of a match due to performance variation.

## 4 Task D

Using our Gibbs samples we would like to estimate the skill difference between Nadal and Federer. We investigate three methods to do this in the following sections.

### 4.1 Approximate marginal skills with Gaussians

We compute the sample marginal mean and variance for each player and model their skills as  $w_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$ . Figure 5 shows the marginal histograms and Gaussian approximation, Table 2 contains the parameters for the approximation. Using the formulas from Task C we find that  $P(w_{Nadal} > w_{Federer}) = 0.438$ . Federer has the higher skill mean with a higher variance than Nadal.

### 4.2 Approximate joint skill distribution with a multivariate Gaussian

We can also approximate the joint skill distribution with a Gaussian by computing the sample covariance matrix, these parameters are again given in Table 2. We find a positive covariance between the two players,  $\sigma_{Nadal, Federer}^2 = 0.009$ . We visualise the joint distribution

$\mu_{Nadal}$	$\mu_{Federer}$	$\sigma_{Nadal}^2$	$\sigma_{Federer}^2$	$\sigma_{Nadal, Federer}^2$
1.48	1.52	0.038	0.041	0.009

Table 2: Parameters for Gaussian fit of marginal and joint skill Gibbs samples for Nadal and Federer

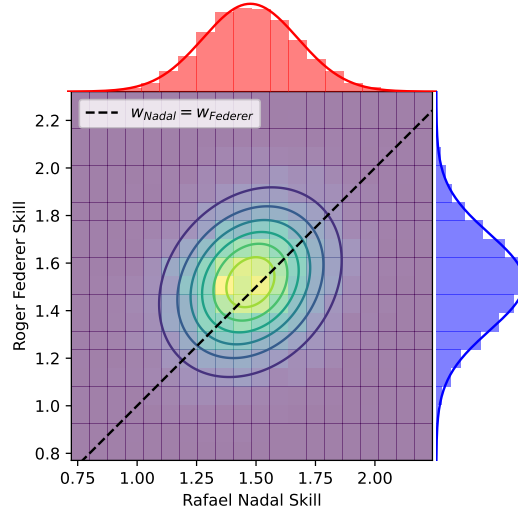


Figure 5: Joint and marginal skill distribution (histograms), joint Gaussian approximation (contours) and marginal gaussian approximation (sides), for Nadal and Federer.

and our approximation in Figure 5, it is clear that the covariance skews the join distribution enough that the marginal approximation is not accurate. We find  $P(s_{ij} > 0) = 0.430$ .

### 4.3 Use the samples directly

We can also compute a Monte Carlo estimate of  $P(w_{Nadal} > w_{Federer})$ . This estimate is given in Equation 4 where  $(w_i^{(n)}, w_j^{(n)})$  are samples from a Markov chain with stationary distribution  $p(w_i, w_j)$ . Our Gibbs samples satisfy this condition and we find  $P(w_{Nadal} > w_{Federer}) = 0.395$ .

$$P(w_i > w_j) = \int \mathbb{1}(w_i > w_j) p(w_i, w_j) dw_i dw_j \approx \frac{1}{N} \sum_{n=1}^N \mathbb{1}(w_i^{(n)} > w_j^{(n)}) \quad (4)$$

### 4.4 Comparison

The skill posterior is not necessarily Gaussian, therefore, computing  $P(w_i > w_j)$  with the Monte Carlo method is the only unbiased method. However, the posterior distribution does look very close to being Gaussian and the Gaussian approximation is a good. Estimating  $P(w_i > w_j)$  with the marginal approximation is far more computationally efficient when comparing a large number of players as we only need to compute a single sum for each player.

$P(s_{ij} > 0)$	Djokovic	Nadal	Federer	Murray
Djokovic	-	0.95	0.92	0.99
Nadal	0.05	-	0.43	0.78
Federer	0.08	0.57	-	0.81
Murray	0.01	0.22	0.19	-

Table 3: Probability that row player is more skilled that column player for top 4 ATP players using Gibbs samples

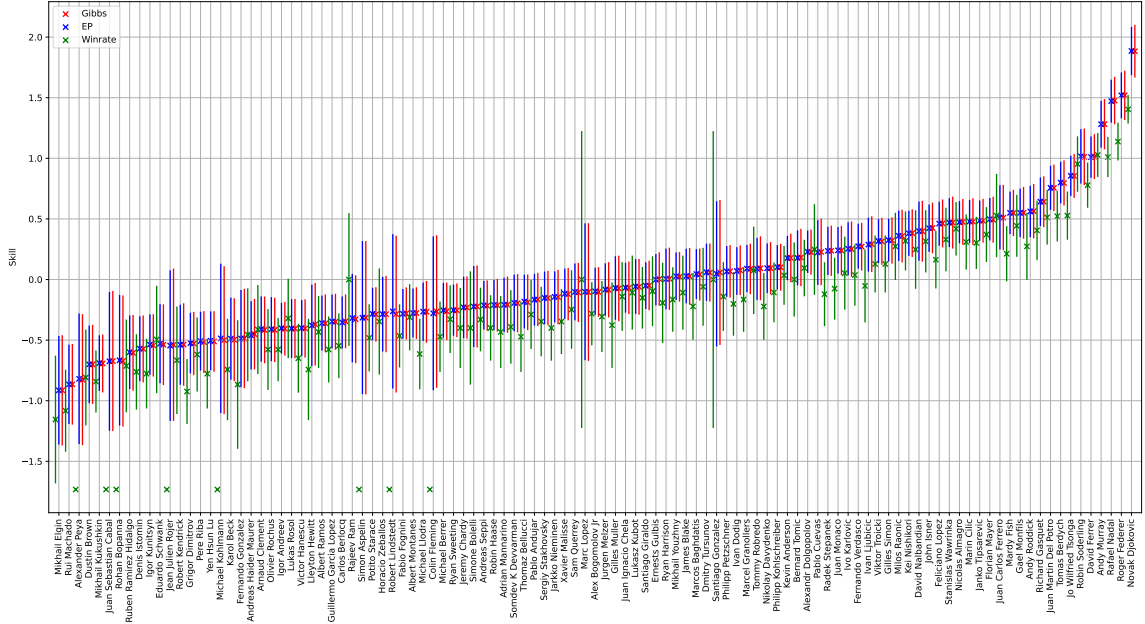


Figure 6: Posterior skill mean and one standard deviation error bars for all players from each model.

We notice that the probabilities in Table 3 are very similar to those in Table 1a. This makes sense as the message passing algorithm approximates the posterior as a Gaussian and the Gibbs samples show this is accurate. The message passing EP algorithm achieves the same performance as marginal Gibbs method but with much faster convergence.

## 5 Task E

We have explored methods for skill estimation using both Gibbs sampling and EP algorithms. A simpler model for estimating skill is to assume every player has a fixed win rate  $r_i$ . The posterior distribution of  $r$  for a player who wins  $k$  out of  $n$  matches with a uniform prior,  $r \sim U(0,1)$ , is the Beta distribution given in Equation 5. We can apply the central limit theorem to approximate this posterior with a Gaussian for large  $n$ . Skill, in this model, is a linear transformation of win rate where  $w_i = \sqrt{12}(r_i - 0.5)$ . This transformation allows gives means the skill prior has zero mean and unit variance for easier model comparison.

$$r|k, n \sim \text{Beta}(1 + k, 1 + n - k) \approx \mathcal{N}\left(\frac{k}{n}, \frac{k(n-k)}{n^3}\right) \quad (5)$$

We show the posterior skill mean and 1 standard deviation error bars for all players from each model in Figure 6. This again shows, as discussed at the end of Task D, that the message passing model achieves similar results to the Gibbs sampler with much faster convergence.

The win rate model ranks the players similarly the other models while having linear time complexity in the number of players. However, there are a few major drawbacks; the model does not take into account the skill of the opponent and the model breaks down for players with few matches. These weaknesses are clear in Figure 6 where some players, who have played few matches and lost them all, have the same minimum possible skill with 0 variance. Andy Murray is an example of a top player who how is ranked differently by the win rate model, being ranked above Roger Federer, this is because on average his matches have been against lower ranked players.