Theory of Complex Systems: Assignment

Henry Zwart (15393879)

1 Modelling the activity of a single neuron

(a) Can you plot the distribution $P(\tau)$ of the time intervals τ between successive spikes? Check that there is indeed a rafactory period, i.e., a time interval τ_0 after each spike, during which the neuron doesn't spike again. What is the duration τ_0 for this time interval?

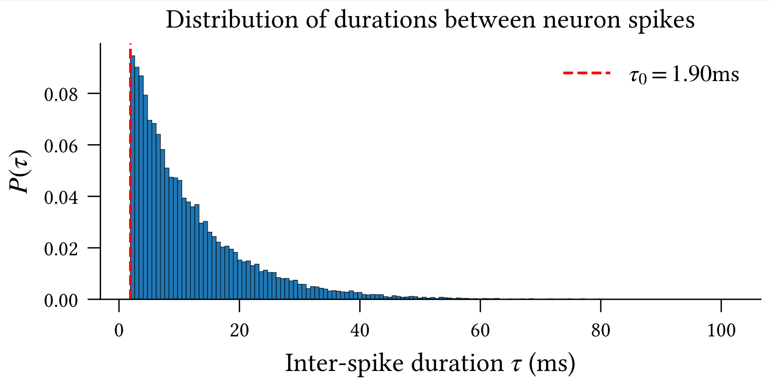


Figure 1: Distribution over the duration between activity spikes for a neuron. A refractory period of $\tau_0 \approx 1.9ms$ is identified as the minimum observed duration between spikes.

Refractory period duration $\tau_0 = 1.9$ms calculated as the minimum observed duration between neuron spikes.

(b) Can you check that the decay of the distribution $P(\tau)$ of inter-spike intervals is indeed exponential? Measure the corresponding decay rate λ

To examine the decay of $P(\tau)$, we consider $\tau > \tau_0 \approx 1.9$. Exponential decay is a good model for $P(\tau - \tau_0)$ if we observe a linear relationship in Equation 1, where the decay rate λ is given by the slope:

$$\begin{aligned}
P(\tau - \tau_0) &= P(\tau_0)e^{-\lambda(\tau-\tau_0)} \\
\frac{1}{P(\tau - \tau_0)} &= \frac{1}{P(\tau_0)}e^{\lambda(\tau-\tau_0)} \\
\log\left(\frac{1}{P(\tau - \tau_0)}\right) &= \log\left(\frac{1}{P(\tau_0)}\right) + \log\left(e^{\lambda(\tau-\tau_0)}\right) \\
-\log(P(\tau - \tau_0)) &= -\log(P(\tau_0)) + \lambda(\tau - \tau_0) \qquad (1)
\end{aligned}$$

To estimate the probability distribution over τ, we subtract the refractory period from each inter-spike duration, and bin the data into 50 bins of uniform width. We normalise the number of observations in each bin with respect to the total number of observations ($n = 30163$) and take this value to be the empirical probability for the value of τ at the center of each bin.

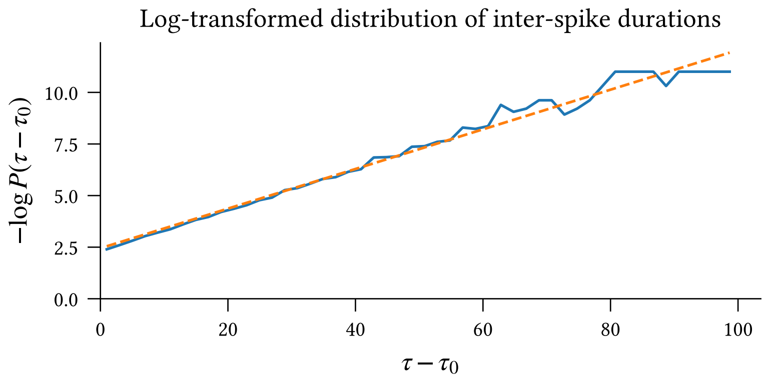


Figure 2: Neuron inter-spike durations (without refractory period) shows a linear relationship under a log-transform, with slope $\lambda = 0.096$ and intercept $P(\tau_0) = 2.4472$. $R = 0.9918$.

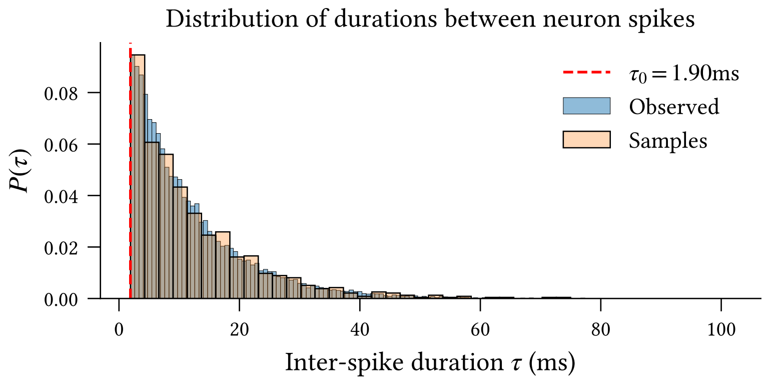
Figure 2 shows the empirical probabilities transformed using Equation 1. We observe a good linear fit, indicating exponential decay in $P(\tau)$ with $\lambda = 0.096$ given by the slope of the linear fit.

The data contains fewer samples for large τ, which explains the reduced quality of the exponential fit.

(c) Can you deduce an analytical expression for the distribution of inter-spike time interval $P(\tau)$ of the delayed Poisson process as a function of λ and τ_0? Compare your model distribution to the one obtained from the data.

$$P(\tau) = \begin{cases} 0 & \tau < \tau_0 \\ \lambda \exp(-\lambda(\tau - \tau_0)) & \tau \geq \tau_0 \end{cases} \tag{2}$$

(d) Using your model, can you generate another 1000 (spike times) datapoints?



(e) What is the average spiking rate f of the neuron in the data? How is f analytically related to τ_0 and λ that you have previously measured?

The average spiking rate, $f = 83.79$hz is calculated as the mean τ^{-1} over the dataset. We can express f analytically in terms of the expected value, as $f = 1/E[\tau]$. First, solving for $E[\tau]$:

$$\begin{aligned}
E[\tau] &= \int_0^\infty \tau P(\tau)\,\mathrm{d}\tau \\
&= \int_0^{\tau_0} \tau P(\tau)\,\mathrm{d}\tau + \int_{\tau_0}^\infty \tau P(\tau)\,\mathrm{d}\tau \\
&= \int_{\tau_0}^\infty \tau \cdot \lambda e^{-\lambda(\tau-\tau_0)}\,\mathrm{d}\tau \\
&= \left[-\tau e^{-\lambda(\tau-\tau_0)}\right]_{\tau_0}^\infty - \int_{\tau_0}^\infty -e^{-\lambda(\tau-\tau_0)}\,\mathrm{d}\tau \\
&= \left[\left[-\tau e^{-\lambda(\tau-\tau_0)}\right] - \left(\frac{1}{\lambda}\right) e^{-\lambda(\tau-\tau_0)}\right]_{\tau_0}^\infty \\
&= \tau_0 + \frac{1}{\lambda}
\end{aligned} \tag{3}$$

The derived value for $E[\tau]$ is as expected, since it represents the expected value of a regular exponential distribution with rate λ, shifted by the refractory period τ_0. The expected spiking rate is then:

$$\begin{aligned}
f &= \frac{1}{E[\tau]} \\
&= \frac{1}{\tau_0 + 1/\lambda} \\
&= \frac{\lambda}{\lambda\tau + 1}
\end{aligned} \tag{4}$$

Using the inferred values for λ and τ_0 from the prior question, Equation 4 gives $f = 81.22$hz.

2 Modelling binary data with the Ising model

2.A Pairwise spin model

(a) How many terms are in the sum over the $\mathsf{pair}(i,j)$? Can you deduce what is the number of parameters in the vector $g = (h_1, ..., h_n, J_{1,2}, ..., J_n$? Can you re-write the sum over the $\mathsf{pair}(i,j)$ as a double sum over i and j (without counting twice each pair)?

The number of pairs in the sum over $\mathsf{pair}(i,j)$ is given by the number of ways which we can choose two distinct spins from a set of n total spins, such that the order of the chosen spins doesn't matter:

$$\frac{n(n-1)}{2} \tag{5}$$

So the total number of parameters in $\boldsymbol{g}$ is:

$$n + \frac{n(n-1)}{2} = \frac{n(n+1)}{2} \tag{6}$$

We can rewrite the sum using a double summation over i and j by enumerating the $k = n-1$ ways to choose the first spin, and then the $n-k$ ways to choose the second:

$$\sum_{\mathsf{pair}(i,j)} J_{i,j} s_i s_j = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} J_{i,j} s_i s_j \tag{7}$$

(b) Can you write down explicitly the terms in the exponential of Eq. (1) for a system with $n = 3$ spins?

$$h_1 s_1 + h_2 s_2 + h_3 s_3 + J_{1,2} s_1 s_2 + J_{1,3} s_1 s_3 + J_{2,3} s_2 s_3 \tag{8}$$

(c) In Eq. (1), we can recognize the Boltzmann distribution, in which the parameter $\beta = 1/(k_b T)$ was taken equal to 1 (more precisely, the constant k_B was taken equal to 1, and the temperature parameter T was absorbed in the parameters h_i and J_{ij}). What is the energy function associated with the Boltzmann distribution in that case? What is the partition function and what is its general expression?

The Boltzmann distribution has the form $P(\hat{\boldsymbol{s}}) = \frac{1}{Z}\exp(-\beta E(\hat{\boldsymbol{s}}))$, where $\beta = \frac{1}{k_b T}$. Then from Eq. (1) we have

$$\begin{aligned} -\frac{1}{k_b T} E(\hat{\boldsymbol{s}}) &= \sum_{i=1}^{n} h_i s_i + \sum_{\text{pair}(i,j)} J_{i,j} s_i s_j \\ \implies \quad E(\hat{\boldsymbol{s}}) &= -k_b T \left[\sum_{i=1}^{n} h_i s_i + \sum_{\text{pair}(i,j)} J_{i,j} s_i s_j \right] \\ &= -\sum_{i=1}^{n} (T \cdot h_i) s_i - \sum_{\text{pair}(i,j)} (T \cdot J_{i,j}) s_i s_j \end{aligned} \tag{9}$$

The partition function Z is the normalisation factor for the Boltzmann distribution:

$$\begin{aligned} \sum_{\hat{\boldsymbol{s}}} \frac{1}{Z} \exp(-\beta E(\hat{\boldsymbol{s}})) &= 1 \\ \implies \quad Z &= \sum_{\hat{\boldsymbol{s}}} \exp(-\beta E(\hat{\boldsymbol{s}})) \end{aligned} \tag{10}$$

For this energy function Z has the form:

$$Z = \sum_{\hat{\boldsymbol{s}}} \exp\left(\sum_{i=1}^{n} h_i s_i + \sum_{\text{pair}(i,j)} J_{i,j} s_i s_j \right) \tag{11}$$

I started trying to derive a closed-form for this, but was taking a while so will come back to it.

(d) Take a spin s_i: if h_i is positive, which direction will s_i tend to turn to, i.e., which direction of s_i will minimize the associated energy $-h_i s_i$? Take a pair of spins s_i and s_j: if J_{ij} is positive, which configurations of (s_i, s_j) minimize the coupling energy $-J_{ij} s_i s_j$?

Assume that we have inferred the best parameters h_i and J_{ij} for the US supreme court dataset discussed in section 2. How would you interpret the sign of the inferred parameters h_i and J_{ij} in this context?

Consider a spin s_i. If h_i is positive, then the component of the spin's energy attributable to h_i is minimised when $s_i > 0$, such that $-h_i s_i < 0$. Similarly, take two spins s_i, s_j with $i \neq j$, then if $J_{ij} > 0$ the energy attributable to the spins' interaction is minimised when $\text{sign}(s_i) = \text{sign}(s_j)$, such that $s_i s_j > 0$ and $-J_{ij} s_i s_j < 0$.

If h_i or J_{ij} were negativethen the opposite result holds ($s_i < 0$, or $\text{sign}(s_i) \neq \text{sign}(s_j)$ respectively). If $h_i = 0$ then s_i has no preferred direction, i.e., the energy due to the field is minimised for any s_i. Likewise, if $J_{ij} = 0$ then any configuration of s_i and s_j minimises their interaction energy.

Suppose now that we have inferred the optimal parameters h_i and J_{ij} for the US supreme court dataset. We can interpret the sign of each parameter by considering its effect *in absence of the other's effect*. For instance, $\text{sign}(h_i)$ signifies the **political leaning** of the i'th judge's votes, in the absence of interactions with other judges. Analogously, $\text{sign}(J_{ij})$ signifies the **tendency for i and j to vote identically**, in the absence of their individual political leanings.

Since we take $+1$ to represent a conservative vote and -1 a liberal one, $h_i < 0$ indicates that i has a tendency to vote liberally, and vice versa. Judges i and j tend to vote similarly if $J_{ij} > 0$, and differently if $J_{ij} < 0$.

Finally, $h_i = 0$ indicates no particular tendency to vote conservative or liberal, and $J_{ij} = 0$ implies no correlation between i and j.

2.B Observables

(a) Given a stationary probability distribution of the state $p_g(\boldsymbol{s})$, what are the definitions of $\langle s_i \rangle$ and of $\langle s_i s_j \rangle$?

For clarity define s_i as a function which extracts the i'th element of a vector $\boldsymbol{s}$, that is $s_i : \hat{\boldsymbol{s}} \mapsto (\hat{\boldsymbol{s}})_i$. Then $\langle s_i \rangle$ is the average local magnetisation of the i'th spin:

$$\langle s_i \rangle = \sum_{\hat{\boldsymbol{s}}} s_i(\hat{\boldsymbol{s}}) \cdot P(\hat{\boldsymbol{s}}) \tag{12}$$

And $\langle s_i s_j \rangle$ is the average local correlation between the i'th and j'th spins:

$$\langle s_i s_j \rangle = \sum_{\hat{\boldsymbol{s}}} s_i(\hat{\boldsymbol{s}}) s_j(\hat{\boldsymbol{s}}) \cdot P(\hat{\boldsymbol{s}}) \tag{13}$$

(b) Consider a dataset $\hat{\boldsymbol{s}}$ composed of N independent observations of the spins: $\hat{\boldsymbol{s}} = \left(\boldsymbol{s}^{(1)}, ..., \boldsymbol{s}^{(N)}\right)$. Let us denote by $\langle s_i \rangle_D$ and $\langle s_i s_j \rangle_D$ the empirical averages of s_i and of $s_i s_j$ respectively (i.e., their average values in the dataset). How would you compute $\langle s_i \rangle_D$ and $\langle s_i s_j \rangle_D$ from the data?

Let D denote the 'dataset', i.e., the set of observation $\left\{\hat{\boldsymbol{s}}^{(1)}, ..., \hat{\boldsymbol{s}}^{(N)}\right\}$. We formalise the notion of the data distribution P_D as the proportion of D comprising observations of a particular microstate:

$$P_D(\hat{\boldsymbol{s}}) = \frac{\#\left\{\hat{\boldsymbol{s}}^{(k)} \in D \mid \hat{\boldsymbol{s}}^{(k)} = \hat{\boldsymbol{s}}\right\}}{N} \tag{14}$$

We may define an estimate for $\langle s_i \rangle$ using P_D, such that it can be computed:

$$\begin{aligned}
\langle s_i \rangle_D &= \sum_{\hat{\boldsymbol{s}}} s_i(\hat{\boldsymbol{s}}) \cdot P_D(\hat{\boldsymbol{s}}) \\
&= \sum_{\hat{\boldsymbol{s}} \notin D} s_i(\hat{\boldsymbol{s}}) \cdot 0 + \sum_{\hat{\boldsymbol{s}} \in D} s_i(\hat{\boldsymbol{s}}) \cdot P_D(\hat{\boldsymbol{s}}) \\
&= \frac{1}{N} \sum_{\hat{\boldsymbol{s}} \in D} s_i(\hat{\boldsymbol{s}}) \cdot \#\left\{\hat{\boldsymbol{s}}^{(k)} \in D \mid \hat{\boldsymbol{s}}^{(k)} = \hat{\boldsymbol{s}}\right\} \\
&= \frac{1}{N} \sum_{k=1}^{N} s_i\left(\hat{\boldsymbol{s}}^{(k)}\right)
\end{aligned} \tag{15}$$

i.e., the empirical average local magnetisation of spin i can be calculated as the average value of s_i as observed in the data. Note that the sum index, $\hat{\boldsymbol{s}} \in D$ on line 3 is taken to mean "$\hat{\boldsymbol{s}}$ occurs in the dataset", rather than being an enumeration of the rows.

Likewise for $\langle s_i s_j \rangle_D$:

$$\begin{aligned}
\langle s_i s_j \rangle_D &= \sum_{\hat{\boldsymbol{s}}} s_i(\hat{\boldsymbol{s}}) s_j(\hat{\boldsymbol{s}}) \cdot P_D(\hat{\boldsymbol{s}}) \\
&= \frac{1}{N} \sum_{\hat{\boldsymbol{s}} \in D} s_i(\hat{\boldsymbol{s}}) s_j(\hat{\boldsymbol{s}}) \cdot \#\left\{\hat{\boldsymbol{s}}^{(k)} \in D \mid \hat{\boldsymbol{s}}^{(k)} = \hat{\boldsymbol{s}}\right\} \\
&= \frac{1}{N} \sum_{k=1}^{N} s_i\left(\hat{\boldsymbol{s}}^{(k)}\right) s_j\left(\hat{\boldsymbol{s}}^{(k)}\right)
\end{aligned} \tag{16}$$

(c) Assume that the data is stationary and that eah datapoint has been randomly sampled from $p(\boldsymbol{s})$. Can you show that the empirical averages, $\langle s_i \rangle_D$ and $\langle s_i s_j \rangle_D$, converge to the model averages, respectively $\langle s_i \rangle$ and $\langle s_i s_j \rangle$, as the number N of datapoints goes to infinity? (very large dataset)

In the previous question we defined $P_D(\hat{\boldsymbol{s}})$, for some microstate $\hat{\boldsymbol{s}}$, as the proportion of observations in the dataset D which were equal to $\hat{\boldsymbol{s}}$. We can equivalently define $P_D(\hat{\boldsymbol{s}})$ as an average over the function $\mathbb{1}\left[\hat{\boldsymbol{s}}^{(k)} = \hat{\boldsymbol{s}}\right]$, which is 1 if this condition holds, and 0 otherwise:

$$P_D(\hat{\boldsymbol{s}}) = \frac{1}{N} \sum_{k=1}^{N} \mathbb{1}\left[\hat{\boldsymbol{s}}^{(k)} = \hat{\boldsymbol{s}}\right] \tag{17}$$

By the Law of Large Numbers, as $n \to \infty$, this average converges to its expected value:

$$\lim_{N \to \infty} P_D(\hat{\boldsymbol{s}}) = P(\hat{\boldsymbol{s}}) \tag{18}$$

The main result follows naturally by considering $\langle s_i \rangle_D$ and $\langle s_i s_j \rangle_D$ as $N \to \infty$:

$$\lim_{N \to \infty} \langle s_i \rangle_D = \lim_{N \to \infty} \left[\sum_{\hat{\boldsymbol{s}} \in D} s_i(\hat{\boldsymbol{s}}) \cdot P_D(\hat{\boldsymbol{s}}) \right] = \sum_{\hat{\boldsymbol{s}}} s_i(\hat{\boldsymbol{s}}) \cdot P(\hat{\boldsymbol{s}}) = \langle s_i \rangle \tag{19}$$

And for the local correlation:

$$\lim_{N \to \infty} \langle s_i s_j \rangle_D = \lim_{N \to \infty} \left[\sum_{\hat{\boldsymbol{s}} \in D} s_i(\hat{\boldsymbol{s}}) s_j(\hat{\boldsymbol{s}}) \cdot P_D(\hat{\boldsymbol{s}}) \right] = \sum_{\hat{\boldsymbol{s}}} s_i(\hat{\boldsymbol{s}}) s_j(\hat{\boldsymbol{s}}) \cdot P(\hat{\boldsymbol{s}}) = \langle s_i s_j \rangle \tag{20}$$

2.C Maximum Entropy models

(a) Consider a spin system with stationary probability distribution $p(\boldsymbol{s})$. Can you recall the definition of the Shannon entropy $S[p(\boldsymbol{s})]$? As mentioned above for the Boltzmann distribution, we will take $k_b = 1$.

With $k_b = 1$, the Shannon entropy is:

$$S[p(\boldsymbol{s})] = -\sum_{\hat{\boldsymbol{s}}} p(\hat{\boldsymbol{s}}) \log p(\hat{\boldsymbol{s}}) \tag{21}$$

The Ising model in Eq. (1) can be seen as a *Maximum Entropy Model*, constrained to reproduce the data local magnetisation and local correlation, i.e., constrained to reproduce all the data averages $\langle s_i \rangle_D$ and $\langle s_i s_j \rangle_D$ (for all spins s_i and s_j). We also want $p(\boldsymbol{s})$ to be normalised, which introduces the additional constraint $\sum_{\boldsymbol{s}} p(\boldsymbol{s}) = 1$. To summarise, we are looking for the set of 2^n probabilities $p(\boldsymbol{s})$ such that $S[p(\boldsymbol{s})]$ is maximal, and such that

$$\sum_{\boldsymbol{s}} p(\boldsymbol{s}) = 1 \quad \text{and} \quad \sum_{\boldsymbol{s}} p(\boldsymbol{s}) s_i(\boldsymbol{s}) = \langle s_i \rangle_D \quad \text{and} \quad \sum_{\boldsymbol{s}} p(\boldsymbol{s}) s_i(\boldsymbol{s}) s_j(\boldsymbol{s}) = \langle s_i s_j \rangle_D \tag{22}$$

where $\langle s_i \rangle_D$ and $\langle s_i s_j \rangle_D$ are constants that are computed from the data for all distinct s_i and s_j. Note that to be more precise, we wrote $s_i(\boldsymbol{s})$ (instead of just s_i) to specify that this is the value of s_i in the state $\boldsymbol{s}$ (this will help with the next questions).

(b) How many constraints are there in total?

The constraints are as follows:

- **Maximisation:** 1 constraint,
- **Normalisation:** 1 constraint,
- **Average local magnetisation:** One per spin (n total)

- **Average local correlation:** One for each pair of distinct spins ($\frac{n(n-1)}{2}$ total)

Thus the total number of constraints is

$$1+1+n+\frac{n(n-1)}{2}=\frac{n(n+1)+1}{2} \tag{23}$$

To find the shape of the distributions $p(\boldsymbol{s})$ that maximises the entropy while satisfying these constraints, we introduce an auxiliary function:

$$\begin{aligned} U[p(\boldsymbol{s})] = S[p(\boldsymbol{s})] + \lambda_0\left(\sum_{\boldsymbol{s}} p(\boldsymbol{s}) - 1\right) + \sum_{i=1}^{n} \alpha_i \left(\sum_{\boldsymbol{s}} p(\boldsymbol{s}) s_i(\boldsymbol{s}) - \langle s_i \rangle_D\right) \\ + \sum_{\mathrm{pair}(i,j)}^{n} \eta_{ij} \left(\sum_{\boldsymbol{s}} p(\boldsymbol{s}) s_i(\boldsymbol{s}) s_j(\boldsymbol{s}) - \langle s_i s_j \rangle_D\right) \end{aligned} \tag{24}$$

where we have introduced a parameter in front of each constraint we want to impose. These parameters (λ_0, α_i, and η_{ij}) are called Lagrange multipliers. To find $p(\boldsymbol{s})$ one must maximise this auxiliary function with respect to the 2^n probabilities $p(\boldsymbol{s})$.

(c) Let us fix a choice of a state $\boldsymbol{s}$. The probability $p_{\boldsymbol{s}} = p(\boldsymbol{s})$ is a parameter of $U[\boldsymbol{p}]$ where $\boldsymbol{p}$ is the vector of the 2^n probabilities. Can you show that:

$$\frac{\partial U[\boldsymbol{p}]}{\partial p_s} = -\log(p_s) - 1 + \lambda_0 + \sum_{i=1}^{n} \alpha_i s_i(\boldsymbol{s}) + \sum_{\mathrm{pair}(i,j)} \eta_{ij} s_i(\boldsymbol{s}) s_j(\boldsymbol{s}) \tag{25}$$

For clarity, we treat the terms in the derivative one at a time. Observe that since we are taking the partial derivative of U with respect to a single element in the vector $p(\boldsymbol{s})$, in each of the sums over $\hat{\boldsymbol{s}}$ in U, all terms will be annhialated by the derivative, with the exception of the particular $\boldsymbol{s}$ which we have fixed.

First, examining the Shannon entropy:

$$\begin{aligned} \frac{\partial}{\partial p_s} S[p(s)] &= \frac{\partial}{\partial p_s}\left[-\sum_{\hat{\boldsymbol{s}}} p(\hat{\boldsymbol{s}}) \log p(\hat{\boldsymbol{s}})\right] \\ &= -\log(p_s) - p_s \cdot \frac{1}{p_s} \\ &= -\log(p_s) - 1 \end{aligned} \tag{26}$$

Next, the normalisation constraint:

$$\frac{\partial}{\partial p_s} \lambda_0 \left(\sum_{\hat{\boldsymbol{s}}} p(\hat{\boldsymbol{s}}) - 1\right) = \frac{\partial}{\partial p_s} \lambda_0 p_s = \lambda_0 \tag{27}$$

The local average magnetisation constraint:

$$\begin{aligned} \frac{\partial}{\partial p_s} \sum_{i=1}^{n} \alpha_i \left(\sum_{\hat{\boldsymbol{s}}} p(\hat{\boldsymbol{s}}) s_i(\hat{\boldsymbol{s}}) - \langle s_i \rangle_D\right) &= \sum_{i=1}^{n} \alpha_i \left(\frac{\partial}{\partial p_s} p_s s_i(\boldsymbol{s})\right) \\ &= \sum_{i=1}^{n} \alpha_i s_i(\boldsymbol{s}) \end{aligned} \tag{28}$$

And lastly the average local correlation constraint:

$$
\begin{aligned}
\frac{\partial}{\partial p_s}\sum_{\text{pair}(i,j)}\eta_{ij}\left(\sum_{\hat{\boldsymbol{s}}} p(\hat{\boldsymbol{s}})s_i(\hat{\boldsymbol{s}})s_j\hat{\boldsymbol{s}} - \langle s_i s_j\rangle_D\right) &= \sum_{\text{pair}(i,j)}\eta_{ij}\left(\frac{\partial}{\partial p_s}p_s s_i(\boldsymbol{s})s_j(\boldsymbol{s})\right)\\
&= \sum_{\text{pair}(i,j)}\eta_{ij}s_i(\boldsymbol{s})s_j(\boldsymbol{s}) \qquad (29)
\end{aligned}
$$

Taking these results together, we arrive at the desired result:

$$
\frac{\partial U[\boldsymbol{p}]}{\partial p_s} = -\log(p_s) - 1 + \lambda_0 + \sum_{i=1}^{n}\alpha_i s_i(\boldsymbol{s}) + \sum_{\text{pair}(i,j)}\eta_{ij}s_i(\boldsymbol{s})s_j(\boldsymbol{s}) \qquad (30)
$$

(d) Can you show that the most general expression of $p_{\boldsymbol{s}}$ with maximal entropy that satisfying the constraints in Eq. (2) is Eq. (1)? Give the relation between λ_0 and the partition function Z. How are the parameters α_i or η_{ij} related to the parameters h_i and J_{ij}?

The constrained optimisation problem described by U is optimised iff, for all microstates $\boldsymbol{s}$ and spins i, j,

$$
\frac{\partial U}{\partial p_s} = \frac{\partial U}{\partial \lambda_0} = \frac{\partial U}{\partial \alpha_i} = \frac{\partial U}{\partial \eta_{ij}} = 0 \qquad (31)
$$

Note that the partial derivatives with respect to the constraints are zero exactly when those constraints are satisfied. The partial derivative with respect to $p_{\boldsymbol{s}}$ is zero for a particular $\boldsymbol{s}$ when

$$
\begin{aligned}
0 &= -\log(p_s) - 1 + \lambda_0 + \sum_{i=1}^{n}\alpha_i s_i(\boldsymbol{s}) + \sum_{\text{pair}(i,j)}\eta_{ij}s_i(\boldsymbol{s})s_j(\boldsymbol{s})\\
p_s &= \exp\left(-(1-\lambda_0) + \sum_{i=1}^{n}\alpha_i s_i(\boldsymbol{s}) + \sum_{\text{pair}(i,j)}\eta_{ij}s_i(\boldsymbol{s})s_j(\boldsymbol{s})\right)\\
&= \frac{1}{\exp(1-\lambda_0)}\exp\left(\sum_{i=1}^{n}\alpha_i s_i(\boldsymbol{s}) + \sum_{\text{pair}(i,j)}\eta_{ij}s_i(\boldsymbol{s})s_j(\boldsymbol{s})\right) \qquad (32)
\end{aligned}
$$

Comparing this expression to Eq. (1), we obtain the following equalities:

$$
\begin{aligned}
Z &= \exp(1-\lambda_0)\\
h_i &= \alpha_i && \text{for each spin } i\\
J_{ij} &= \eta_{ij} && \text{for each pair of spins } i, j \qquad (33)
\end{aligned}
$$

2.D Statistical inference: model with no couplings

Consider the model with no couplings (all the $J_{ij} = 0$):

$$
p_{\boldsymbol{g}}(\boldsymbol{s}) = \frac{1}{Z(\boldsymbol{g})}\exp\left(\sum_{i=1}^{n}h_i s_i\right) \qquad (34)
$$

The vector $\boldsymbol{g} = (h_1, ..., h_n)$ now only contains n local field parameters.

(a) Can you show that in that case the model is assuming the variables are independent from each other, i.e., that we can write the joint probability distribution as a product of a probability distribution over each variable $p_{\boldsymbol{g}}(\boldsymbol{s}) = \prod_{i=1}^{n} p_{\boldsymbol{h}_i}(s_i)$? What is the probability distribution $p_{\boldsymbol{h}_i}(s_i)$ for the spin variable s_i?

We first solve for the partition function in the described model. Using the normalisation condition, we have:

$$\begin{aligned}
Z(\boldsymbol{g}) &= \sum_{\hat{\boldsymbol{s}}} \exp\left(\sum_{i=1}^{n} h_i s_i\right) \\
&= \sum_{s_1 \in \pm 1} \cdots \sum_{s_{n-1} \in \pm 1} \exp\left(\sum_{i=1}^{n-1} h_i s_i\right)(\exp(h_n) + \exp(-h_n)) \\
&= \prod_{i=1}^{n} \exp(h_i) + \exp(-h_i) \\
&= \prod_{i=1}^{n} 2\cosh(h_i) \qquad (35)
\end{aligned}$$

We can then rewrite $p_{\boldsymbol{g}}(\boldsymbol{s})$ as:

$$\begin{aligned}
p_{\boldsymbol{g}}(\boldsymbol{s}) &= \frac{1}{Z(\boldsymbol{g})} \exp\left(\sum_{i=1}^{n} h_i s_i\right) \\
&= \frac{1}{Z(\boldsymbol{g})} \prod_{i=1}^{n} \exp(h_i s_i) \\
&= \prod_{i=1}^{n} \frac{\exp(h_i s_i)}{2\cosh(h_i)} \qquad (36)
\end{aligned}$$

Where we recognise the term inside the product as the probability distribution for a system consisting of a single spin. It follows that $\frac{1}{2\cosh(h_i)}$ normalises the distribution for a single spin, and thus we find $p_{\boldsymbol{g}}(\boldsymbol{s}) = \prod_{i=1}^{n} p_{\boldsymbol{h_i}}(s_i)$, where:

$$p_{\boldsymbol{h_i}}(s_i) = \frac{\exp(h_i s_i)}{2\cosh(h_i)} \qquad (37)$$

(b) Take one of the spin variables s_i. We recall that $\langle s_i \rangle_D$ is the average value of s_i in the data (given a dataset, this quantity is a constant), and that $\langle s_i \rangle = \sum_{\boldsymbol{s}} p(\boldsymbol{s}) s_i$ is the model average of s_i. Can you show that the value of the parameter h_i that satisfies the constraint $\langle s_i \rangle = \langle s_i \rangle_D$ is:

$$h_i = \tanh^{-1}(\langle s_i \rangle_D), \qquad (38)$$

where $\tanh^{-1}(x)$ denotes the inverse of the hyperbolic tangent? In particular, in that case the probability distribution over s_i in the model is exactly equal to the empirical distribution of s_i.

Let i be a spin in a model with n spins, and $\langle s_i \rangle_D$ the average value of s_i in a dataset D. Suppose that $\langle s_i \rangle = \langle s_i \rangle_D$ in the model, then from the definition of $\langle s_i \rangle$

$$\langle s_i \rangle = \sum_{\hat{\boldsymbol{s}}} s_i(\hat{\boldsymbol{s}}) \cdot p_{\boldsymbol{g}}(\hat{\boldsymbol{s}}) = \langle s_i \rangle_D \qquad (39)$$

From the previous exercise, we have that $p_{\boldsymbol{g}}(\hat{\boldsymbol{s}}) = \prod_{j=1}^{n} p_{\boldsymbol{h_j}}(s_j)$. Taking $\hat{\boldsymbol{s}}_i = (s_1, ..., s_{i-1}, s_{i+1}, ..., s_n)$ to denote the vector of spin variables excluding s_i, we use this result to rewrite Equation 39 in terms of $p_{\boldsymbol{h_i}}$:

$$
\begin{aligned}
\langle s_i \rangle_D &= \sum_{\hat{\boldsymbol{s}}} s_i(\hat{\boldsymbol{s}}) \cdot p_{\boldsymbol{h}_i}(s_i(\hat{\boldsymbol{s}})) \cdot \prod_{j \neq i} p_{\boldsymbol{h}_j}(s_j(\hat{\boldsymbol{s}})) \\
&= \sum_{s \in \pm 1} s p_{\boldsymbol{h}_i}(s) \cdot \overbrace{\sum_{\hat{\boldsymbol{s}}_{-i}} \prod_{j \neq i} p_{\boldsymbol{h}_j(\boldsymbol{s}_j(\hat{\boldsymbol{s}}_{-i}))}}^{=1} \\
&= p_{\boldsymbol{h}_i}(1) - p_{\boldsymbol{h}_i}(-1) \\
&= \frac{\exp(h_i) - \exp(-h_i)}{2\cosh(h_i)} \\
&= \frac{\sinh(h_i)}{\cosh(h_i)} \\
&= \tanh(h_i) \qquad (40)
\end{aligned}
$$

In the second line of reasoning we group the original summation terms by their s_i component. The inner summation is then the sum over the probabilities of observing any other microstate for the remaining $n-1$ variables. The normalisation condition of probability distributions implies that this is equal to 1.

Finally, we obtain the desired result, $h_i = \tanh^{-1}(\langle s_i \rangle_D)$.

(c) In Eq. (6), we observe that:

- If $\langle s_i \rangle_D > 0$, then the inferred h_i is also positive;
- Reciprocally, $\langle s_i \rangle_D < 0$, then the inferred h_i is also negative.

How does this connect with the tendency of the i'th judge to vote on average more liberal or more conservative? Is this result coherent with the general comments that we did in Question Q1.4.?

This result is coherent with our comments in Question Q1.4, in which we interpreted the sign of h_i as reflecting the sign of i's political leaning (-1 for liberal, $+1$ for conservative). As $\langle s_i \rangle_D$ is the average value of s_i in a dataset D, a positive value occurs when i votes conservative in more than 50% of cases. Likewise, a negative value occurs when i votes liberal in more than 50% of cases.

As in Q1.4, if the i'th judge has an equal number of liberal and conservative votes in D, then $\langle s_i \rangle_D = h_i = 0$.

While in Q1.4 our discussion was concerned with a model which included interactions, the comments are still relevant here, as we interpreted the sign of h_i to reflect political leaning in absence of interactions with other judges. In this model we simply make this assumption explicit by taking $J_{ij} = 0$.

2.E Statistical inference: maximising the log-likelihood function

Introducing the likelihood function. Looking more closely at Eq. (1), one can see that it does not just define a single probability distribution, but many of them: there is one probability distribution for each value of the set of parameters $\boldsymbol{g}$. More precisely, the distribution in Eq. (1) changes continuously as one continuously varies the parameters in $\boldsymbol{g}$. We say that Eq. (1) defines a *parametric family of probability distributions*. The inference procedure consists in finding the value of the parameters $\boldsymbol{g}$ that maximises the probability that the model $p_{\boldsymbol{g}}(\boldsymbol{s})$ produces the data.

To do so, we introduce the *log-likelihood function*:

$$\mathcal{L}(\boldsymbol{g}) = \log P_{\boldsymbol{g}}(\hat{\boldsymbol{s}}) \qquad (41)$$

where $P_{\boldsymbol{g}}(\hat{\boldsymbol{s}})$ is the probability that the model $p_{\boldsymbol{g}}(\boldsymbol{s})$ produces the dataset $\hat{\boldsymbol{s}} = \left(\boldsymbol{s}^{(1)}, ..., \boldsymbol{s}^{(N)}\right)$. Note that $\mathcal{L}(\boldsymbol{g})$ is a function of the parameters $\boldsymbol{g}$. The inference procedure therefore consists in finding the value $\boldsymbol{g}^\star$ of the parameters that maximises $\mathcal{L}(\boldsymbol{g})$. For the moment we will assume that there exists only a unique such value of $\boldsymbol{g}$.

(a) We assume that, in the dataset $\hat{\boldsymbol{s}}$, all datapoints are independently samples from an underlying distribution $p_{\boldsymbol{g}}(\boldsymbol{s})$. Can you show that the log-likelihood function can be re-written as: $\mathcal{L} = N \sum_{\boldsymbol{s}} p_D(\boldsymbol{s}) \log p_{\boldsymbol{g}}(\boldsymbol{s})$ where $p_D(\boldsymbol{s})$ is the empirical distribution over the states? The empirical distribution is given by $p_D(\boldsymbol{s}) = \frac{K(\boldsymbol{s})}{N}$ where $K(\boldsymbol{s})$ is the number of times that the datapoint $\boldsymbol{s}$ occurs in the dataset.

From the independent sampling assumption, we have that $P_{\boldsymbol{g}}(\hat{\boldsymbol{s}}) = \prod_{k=1}^N p_{\boldsymbol{g}}(\hat{\boldsymbol{s}}^{(k)})$. Substituting this in Equation 41, we have:

$$\begin{aligned}
\mathcal{L}(\boldsymbol{g}) &= \log\left(\prod_{k=1}^{N} p_{\boldsymbol{g}}\big(\hat{\boldsymbol{s}}^{(k)}\big)\right) \\
&= \sum_{k=1}^{N} \log p_{\boldsymbol{g}}\big(\hat{\boldsymbol{s}}^{(k)}\big) \\
&= \sum_{\boldsymbol{s}} K(\boldsymbol{s}) \log p_{\boldsymbol{g}}(\boldsymbol{s}) \\
&= \sum_{\boldsymbol{s}} N \cdot \frac{K(\boldsymbol{s})}{N} \log p_{\boldsymbol{g}}(\boldsymbol{s}) \\
&= N \sum_{\boldsymbol{s}} p_D(\boldsymbol{s}) \log p_{\boldsymbol{g}}(\boldsymbol{s}) \qquad (42)
\end{aligned}$$

Ising model. We now take the model distribution $p_{\boldsymbol{g}}(\boldsymbol{s})$ to be given by the Ising model in Eq. (1).

(b) Taking the first derivative of $\mathcal{L}(\boldsymbol{g})$ with respect to a parameter h_i, can you show that at the maximum of $\mathcal{L}(\boldsymbol{g})$ we have that $\langle s_i \rangle = \langle s_i \rangle_D$? Similarly, taking the first derivative of $\mathcal{L}(\boldsymbol{g})$ with respect to a parameter J_{ij}, can you show that at the maximum of $\mathcal{L}(\boldsymbol{g})$ we have that $\langle s_i s_j \rangle = \langle s_i s_j \rangle_D$?

Let $p_{\boldsymbol{g}}(\boldsymbol{s})$ be as defined in Eq. (1), and let i correspond to the spin s_i. Then $\mathcal{L}(\boldsymbol{g})$ is:

$$\begin{aligned}
\mathcal{L}(\boldsymbol{g}) &= N \sum_{\boldsymbol{s}} p_D(\boldsymbol{s}) \log\left[\frac{1}{Z(\boldsymbol{g})} \exp\left(\sum_{j=1}^{n} h_j s_j + \sum_{\text{pair}(j,k)} J_{jk} s_j s_k\right)\right] \\
&= N \sum_{\boldsymbol{s}} p_D(\boldsymbol{s}) \left[\sum_{j=1}^{n} h_j s_j + \sum_{\text{pair}(j,k)} J_{jk} s_j s_k - \log Z(\boldsymbol{g})\right] \qquad (43)
\end{aligned}$$

The partial derivative of $\mathcal{L}(\boldsymbol{g})$ with respect to h_i is then:

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial h_i} &= N \sum_{\boldsymbol{s}} p_D(\boldsymbol{s}) \left[\frac{\partial}{\partial h_i} \sum_{j=1}^{n} h_j s_j + \frac{\partial}{\partial h_i} \sum_{\text{pair}(j,k)} J_{jk} s_j s_k - \frac{\partial}{\partial h_i} \log Z(\boldsymbol{g})\right] \\
&= N \sum_{\boldsymbol{s}} p_D(\boldsymbol{s}) \cdot \left(s_i - \frac{\partial}{\partial h_i} \log Z(\boldsymbol{g})\right) \qquad (44)
\end{aligned}$$

Where the partial derivative of the log partition function with respect to h_i is:

$$\begin{aligned}
\frac{\partial \log Z(\boldsymbol{g})}{\partial h_i} &= \frac{1}{Z(\boldsymbol{g})} \cdot \frac{\partial}{\partial h_i} \sum_{\boldsymbol{s}'} \exp\left(\sum_{j=1}^{n} h_j s_j + \sum_{\text{pair}(j,k)} J_{jk} s_j s_k\right) \\
&= \frac{1}{Z(\boldsymbol{g})} \cdot \sum_{\boldsymbol{s}'} \frac{\partial}{\partial h_i} \exp\left(\sum_{j=1}^{n} h_j s_j + \sum_{\text{pair}(j,k)} J_{jk} s_j s_k\right) \\
&= \frac{1}{Z(\boldsymbol{g})} \cdot \sum_{\boldsymbol{s}'} \exp\left(\sum_{j=1}^{n} h_j s_j + \sum_{\text{pair}(j,k)} J_{jk} s_j s_k\right) \cdot s_i \\
&= \sum_{\boldsymbol{s}'} p_{\boldsymbol{g}}(\boldsymbol{s}') s_i \\
&= \langle s_i \rangle \qquad (45)
\end{aligned}$$

$\mathcal{L}(\boldsymbol{g})$ attains its maximum value with respect to h_i when $\frac{\partial \mathcal{L}}{\partial h_i} = 0$. Substituting Equation 45 into Equation 44 and solving for the maximum, we find:

$$
\begin{aligned}
0 &= N\sum_{\boldsymbol{s}} p_D(\boldsymbol{s})\cdot\left(s_i - \langle s_i\rangle\right)\\
\implies \qquad N\sum_{\boldsymbol{s}} p_D(\boldsymbol{s})\langle s_i\rangle &= N\sum_{\boldsymbol{s}} p_D(\boldsymbol{s}) s_i\\
\langle s_i\rangle\cdot\sum_{\boldsymbol{s}} p_D(\boldsymbol{s}) &= \langle s_i\rangle_D\\
\langle s_i\rangle &= \langle s_i\rangle_D \qquad (46)
\end{aligned}
$$

Where the left-hand side summation cancels in the final step since the occurrence proportions of states in D must sum to 1.

We proceed analogously to show that $\langle s_i s_j\rangle = \langle s_i s_j\rangle_D$ when $\mathcal{L}(\boldsymbol{g})$ is maximised with respect to J_{ij}:

$$
\begin{aligned}
\frac{\partial\mathcal{L}}{\partial J_{ij}} &= N\sum_{\boldsymbol{s}} p_D(\boldsymbol{s})\left[\frac{\partial}{\partial J_{ij}}\sum_{k=1}^{n} h_k s_k + \frac{\partial}{\partial J_{ij}}\sum_{\text{pair}(k,\ell)} J_{k\ell}s_k s_\ell - \frac{\partial}{\partial J_{ij}}\log Z(\boldsymbol{g})\right]\\
&= N\sum_{\boldsymbol{s}} p_D(\boldsymbol{s})\cdot\left(s_i - \frac{\partial}{\partial h_i}\log Z(\boldsymbol{g})\right)\\
&= N\sum_{\boldsymbol{s}} p_D(\boldsymbol{s})\cdot\left(s_i - \sum_{\boldsymbol{s}'} p_{\boldsymbol{g}}(\boldsymbol{s}') s_i s_j\right)\\
&= N\sum_{\boldsymbol{s}} p_D(\boldsymbol{s})\cdot\left(s_i - \langle s_i s_j\rangle\right) \qquad (47)
\end{aligned}
$$

As before, evaluating Equation 47 at $\frac{\partial\mathcal{L}}{\partial J_{ij}} = 0$ gives:

$$
\begin{aligned}
N\sum_{\boldsymbol{s}} p_D(\boldsymbol{s})\langle s_i s_j\rangle &= N\sum_{\boldsymbol{s}} p_D(\boldsymbol{s}) s_i s_j\\
\langle s_i s_j\rangle &= \langle s_i s_j\rangle_D \qquad (48)
\end{aligned}
$$

3 Application to the analysis of the US supreme Court

A system with n spin variables can be in 2^n different states. However, most of the time, the number of different states observed in a dataset is very small compared to 2^n.

(a) For the US Supreme court dataset: What is the number n of spin variables, and the total number 2^n of states that can be observed for that system? What is the total number N of datapoints in the dataset? What is the number N_{max} of different states that are observed?

(b)

(c)

(d)

(e)

(f)

(g)

(h)

(i)