Theory of Complex Systems: Assignment

Henry Zwart (15393879)

1 Modelling the activity of a single neuron

Q1. 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 shows the distribution of inter-spike durations in the dataset as a histogram with 50 bins of uniform width. We obseve a refractory period of $\tau_0 = 1.9$ms, indicated by the lack of inter-spike durations at the left-hand side of the distribution, in the interval $[0, \tau_0)$.

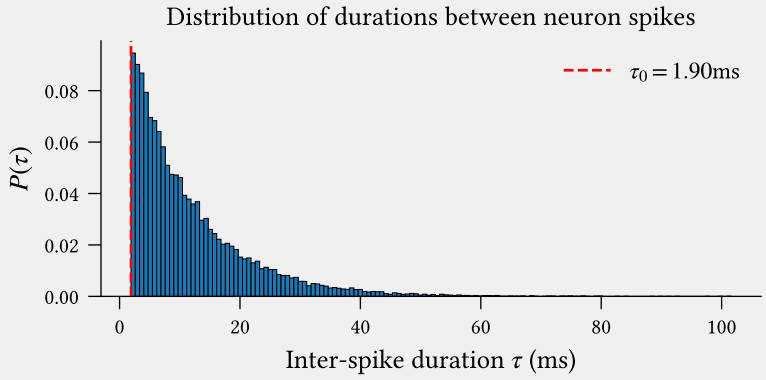


Figure 1: Distribution over the duration between activity spikes for a neuron. A refractory period of $\tau_0 \approx 1.9ms$ is identified as the minimum observed duration between spikes.

Q2. Can you check that the decay of the distribution $P(\tau)$ of inter-spike intervals is indeed exponential? Measure the corresponding decay rate λ

We say $P(\tau)$ decreases with exponential decay for increasing $\tau > \tau_0$ if it is well-described by a model of the form

$$P(\tau - \tau_0) = P(\tau_0)e^{-\lambda(\tau - \tau_0)} \tag{1}$$

To test this hypothesis, we can check whether the observed inter-spike durations display a good linear fit under the equivalent linear model:

$$\begin{aligned} P(\tau - \tau_0) = &\quad P(\tau_0)e^{-\lambda(\tau-\tau_0)} \\ \log P(\tau - \tau_0) = &\quad \log P(\tau_0) - \lambda(\tau - \tau_0) \\ -\log P(\tau - \tau_0) = &-\log P(\tau_0) + \lambda(\tau - \tau_0) \end{aligned} \tag{2}$$

We estimate $P(\tau - \tau_0)$ by subtracting the calculated refractory period from each observed inter-spike duration, then binning the data into 50 bins of uniform width. We normalise the observation count in each bin with respect to the total number of observations ($n = 30163$), and take this to be the empirical probability for the value of τ at the center of each bin.

Figure 2 shows the empirical probabilities transformed using Equation 2. 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 quality of the linear fit is reduced for large τ; however, this is explained by a smaller sample size in this range.

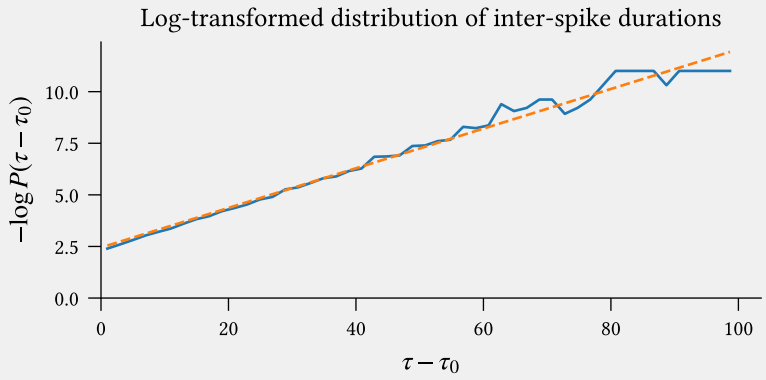


Figure 2: Neuron inter-spike durations (without refractory period) shows a linear relationship under a log-transform, with slope $\lambda = 0.096$ and intercept $P(\tau_0) = 2.4472$. $R = 0.9918$.

Q3. 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.

In the previous question we found that $P(\tau)$ exhibited exponential decay for $\tau > \tau_0$. As such, we can model $P(\tau)$ as a shifted exponential distribution where $P(\tau) = 0$ for $\tau < \tau_0$, and the rate λ is as calculated in **Q2**.

$$P(\tau) = \begin{cases} 0 & \tau < \tau_0 \\ \lambda \exp(-\lambda(\tau - \tau_0)) & \tau \geq \tau_0 \end{cases} \tag{3}$$

For Equation 3 to be a probability distribution it remains to show that it is well-normalised. To see that this is true, observe that the probability density for $\tau < \tau_0$ is 0, and the density for $\tau \geq \tau_0$ is 1 (since this is just a standard exponential distribution). Thus it follows that

$$\begin{aligned} \int_0^\infty P(t)\,\mathrm{d}t &= \cancel{\int_0^{\tau_0} P(t)\,\mathrm{d}t} + \int_{\tau_0}^\infty \lambda e^{-\lambda(t-\tau_0)}\,\mathrm{d}t \\ &= \int_0^\infty \lambda e^{-\lambda t}\,\mathrm{d}t \\ &= 1 \end{aligned} \tag{4}$$

And so Equation 3 is an analytical expression for the distribution $P(\tau)$.

Q4. Using your model, can you generate another 1000 (spike times) datapoints?

To sample 1000 new spike times, it suffices to sample 1000 inter-spike durations $\boldsymbol{\tau} = (\tau_{n+1}, ..., \tau_{n+1000})$ from Equation 3, and take the spike times $\boldsymbol{t} = (t_{n+1}, ..., t_{n+1000})$ with $t_{n+k} = t_n + \sum_{i=1}^k \tau_{n+i}$ (i.e., cumulative sum of interspike durations to t_{n+k}).

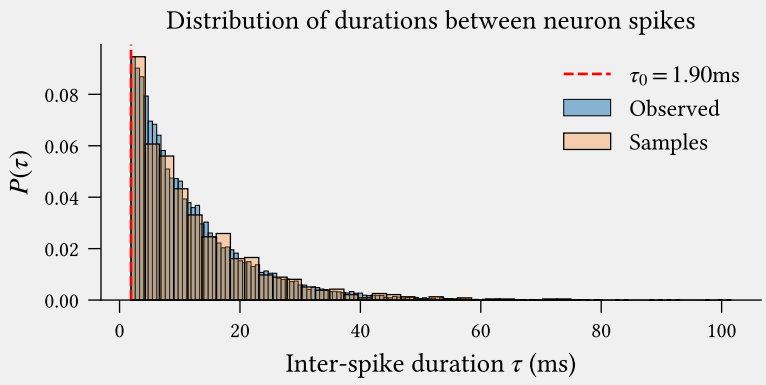
However, the task of sampling from Equation 3 remains. We do this via the inverse-transform method. Letting $F(T)$ be the cumulative distribution associated with $P(\tau)$, we sample $u \in \mathcal{U}_{[0,1]}$ and calculate $\tau = F^{-1}(u)$ (so long as F is invertible). Then each τ is sampled with probability equal to $P(\tau)$. $F(T)$ is derived as follows:

$$
\begin{aligned}
F(T) &= \int_0^T P(t)\,\mathrm{d}t \\
&= \int_{\tau_0}^T P(t)\,\mathrm{d}t \\
&= \left[-\exp(-\lambda(t-\tau_0))\right]_{\tau_0}^T \\
&= \begin{cases} 0 & \text{if } T < \tau_0 \\ 1 - e^{-\lambda(T-\tau_0)} & \text{if } T \geq \tau_0 \end{cases}
\end{aligned} \tag{5}
$$

While $F(T)$ is piecewise this is not a problem for the inverse-transform since (in practice) $u \neq 0$. Indeed Equation 5 is invertible for $u \in (0, 1)$, with:

$$
\begin{aligned}
u &= 1 - e^{-\lambda(t-\tau_0)} \\
1 - u &= e^{-\lambda(t-\tau_0)} \\
-\lambda(t-\tau_0) &= \ln(1-u) \\
t &= -\frac{1}{\lambda}\ln(1-u) + \tau_0
\end{aligned} \tag{6}
$$

Figure 3 shows the distribution over additional samples $\boldsymbol{\tau} = (\tau_{n+1}, ..., \tau_{n+1000})$ as sampled using the inverse-transform method with Equation 6, overlaid on the distribution of observed data.



Q5. 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 is defined as the inverse of the expected duration between spikes, $E[\tau]$. To determine the analytical form of f, we first derive $E[\tau]$:

$$
\begin{aligned}
E[\tau] &= \cancel{\int_{\tau_0}^{\tau_0} tP(t)\,\mathrm{d}t} + \int_{\tau_0}^{\infty} t \cdot \lambda e^{-\lambda(t-\tau_0)}\,\mathrm{d}t \\
&= \left[-te^{-\lambda(t-\tau_0)}\right]_{\tau_0}^{\infty} - \int_{\tau_0}^{\infty} -e^{-\lambda(t-\tau_0)}\,\mathrm{d}t \qquad \text{(integration by parts)} \\
&= \left[-te^{-\lambda(t-\tau_0)} - \left(\frac{1}{\lambda}\right)e^{-\lambda(t-\tau_0)}\right]_{\tau_0}^{\infty} \\
&= \tau_0 + \frac{1}{\lambda}
\end{aligned} \tag{7}
$$

We then derive f as $\frac{1}{E[\tau]}$:

$$f = 1/E[\tau] = \frac{1}{\tau_0 + \frac{1}{\lambda}} \implies \frac{\lambda}{\lambda\tau_0 + 1} \tag{8}$$

Evaluating Equation 8 using our estimates for τ_0 (**Q1.**) and λ (**Q2.**), we obtain an estimate for the average spiking rate $f = 81.22$hz. We can compare this to the average spiking rate as calculated directly from that data as the average value of τ^{-1}, $f_{\text{data}} = 83.79$hz. The difference between the two values is likely the result of numerical error differences between the two methods.

2 Modelling binary data with the Ising model

2.1 Pairwise spin model

It is common to model the collective behavior of systems of binary variables with Ising-like models. To do so, we assume that the system is in a stationary state, and therefore that the datapoints are independently sampled from the same stationary probability distribution. We take this probability distribution to have the general form of an Ising model:

$$p_{\boldsymbol{g}}(\boldsymbol{s}) = \frac{1}{Z(\boldsymbol{g})} \exp\left(\sum_{i=1}^{n} h_i s_i + \sum_{\text{pair}(i,j)} J_{ij} s_i s_j\right) \tag{9}$$

where n is the number of spin variables, $\text{pair}(i,j)$ denotes a summation over all possible pairs of distinct spin variables, $\boldsymbol{g} = \left(h_1, ..., h_n, J_{1,2}, ..., J_{n-1,n}\right)$ is a vector of (real) parameters, and $Z(\boldsymbol{g})$ is a normalisation factor. There are several differences compared to the Ising model we have seen in class:

- There is a different external field h_i for each spin s_i, which can take any real value. In particular, the h_i's are not necessarily all positive or all negative.
- There is a different coupling parameter J_{ij} for each pair of spins s_i and s_j. The parameter J_{ij} parametrises the strength of the coupling between s_i and s_j, and can take any real value. In particular, the J_{ij}'s are not necessarily all positive or all negative.
- The summation is over all possible pairs (i,j) of spins, and not just over the "nearest neighbors". The reason is that, in a general dataset, we have a priori no idea if there exists an underlying structure between the variables and if so, what that structure is, and therefore we don't know which variables are "nearest neighbors".

The general goal of the problem is to infer the set of parameters $\boldsymbol{g} = \left(h_1, ..., h_n, J_{1,2}, ...J_{n-1,n}\right)$ that is the most appropriate to model the data, i.e. to find the parameters $\boldsymbol{g}$ for which the probability distribution in Equation 9 best fits the data. This way, we would infer from the data the underlying structure between the variables, and find which variables tend to be strongly influenced by an external parameter, and which pairs of variables tend to be more strongly coupled.

Q1.1. How many terms are in the sum over the $\text{pair}(i,j)$? Can you deduce what is the number of parameters in the vector $g = \left(h_1, ..., h_n, J_{1,2}, ..., J_n\right)$? Can you re-write the sum over the $\text{pair}(i,j)$ as a double sum over i and j (without counting twice each pair)?

Assuming that $\text{pair}(i,j)$ is order-independent, the number of pairs in the sum over $\text{pair}(i,j)$ 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: $n(n-1)/2$. So the total number of parameters in $\boldsymbol{g}$ is:

$$n + \frac{n(n-1)}{2} = \frac{n(n+1)}{2} \tag{10}$$

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 (given the first):

$$\sum_{\text{pair}(i,j)} J_{i,j} s_i s_j = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} J_{i,j} s_i s_j \tag{11}$$

Q1.2. Can you write down explicitly the terms in the exponential of Equation 9 for a system with $n = 3$ spins?

$$h_1 s_1 + h_2 s_2 + h_3 s_3 + J_{1,2} s_1 s_2 + J_{1,3} s_1 s_3 + J_{2,3} s_2 s_3 \tag{12}$$

Q1.3. In Equation 9, 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 Equation 9 we have

$$\begin{aligned} -\frac{1}{k_b T} E(\hat{\boldsymbol{s}}) &= \sum_{i=1}^{n} h'_i s_i + \sum_{\text{pair}(i,j)} J'_{i,j} s_i s_j \\ \implies \quad E(\hat{\boldsymbol{s}}) &= -k_b T \left[\sum_{i=1}^{n} h'_i s_i + \sum_{\text{pair}(i,j)} J'_{i,j} s_i s_j \right] \\ &= -\sum_{i=1}^{n} (T \cdot h'_i) s_i - \sum_{\text{pair}(i,j)} \big(T \cdot J'_{i,j}\big) s_i s_j \end{aligned} \tag{13}$$

As stated, the temperature parameter T is absorbed into the parameters h_i and J_{ij} – that is, for constant temperature T, we redefine $h_i = T \cdot h'_i$ and $J_{ij} = T \cdot J'_{ij}$, giving us the following energy function:

$$E(\hat{\boldsymbol{s}}) = -\sum_{i=1}^{n} h_i s_i - \sum_{\text{pair}(i,j)} J_{ij} s_i s_j \tag{14}$$

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

$$\sum_{\hat{\boldsymbol{s}}} \frac{1}{Z(\boldsymbol{g})} \exp(-\beta E(\hat{\boldsymbol{s}})) = 1 \qquad \implies \qquad Z(\boldsymbol{g}) = \sum_{\hat{\boldsymbol{s}}} \exp(-\beta E(\hat{\boldsymbol{s}})) \tag{15}$$

For the energy function in Equation 14 $Z(\boldsymbol{g})$ has the form:

$$Z(\boldsymbol{g}) = \sum_{\hat{\boldsymbol{s}}} \exp\left(\sum_{i=1}^{n} h_i s_i + \sum_{\text{pair}(i,j)} J_{ij} s_i s_j \right) \tag{16}$$

Q1.4. 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 negative then 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, aside from that which may arise from nonzero h_i, h_j.

2.2 Observables

The important observables of the system are the local average magnetisations $\langle s_i \rangle$ and the local correlations $c_{ij} = \langle s_i s_j \rangle - \langle s_i \rangle \langle s_j \rangle$, where the angle brackets $\langle A(\boldsymbol{s}) \rangle$ denote the ensemble average (or thermal average) of the microscopic quantity $A(\boldsymbol{s})$.

Q2.1. 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 s, that is $s_i : \boldsymbol{s} \mapsto (\boldsymbol{s})_i$. Then $\langle s_i \rangle$ is the average local magnetisation of the i'th spin:

$$\langle s_i \rangle = \sum_{\boldsymbol{s}} s_i(\boldsymbol{s}) \cdot p_{\boldsymbol{g}}(\boldsymbol{s}) \tag{17}$$

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

$$\langle s_i s_j \rangle = \sum_{\boldsymbol{s}} s_i(\boldsymbol{s}) s_j(\boldsymbol{s}) \cdot p_{\boldsymbol{g}}(\boldsymbol{s}) \tag{18}$$

Q2.2. 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 $\hat{\boldsymbol{s}} = \left(\boldsymbol{s}^{(1)}, ..., \boldsymbol{s}^{(N)}\right)$ be a dataset of observed microstates. We formalise the notion of the data distribution $p_{\hat{\boldsymbol{s}}}$ as the proportion of $\hat{\boldsymbol{s}}$ comprising observations of a particular microstate:

$$p_{\hat{\boldsymbol{s}}}(\boldsymbol{s}) = \frac{\#\left\{\boldsymbol{s}^{(k)} \in \hat{\boldsymbol{s}} \mid \boldsymbol{s}^{(k)} = \boldsymbol{s}\right\}}{N} \tag{19}$$

We define $\langle s_i \rangle_D$ using $p_{\hat{\boldsymbol{s}}}$, in such a form that it can be computed from $\hat{\boldsymbol{s}}$:

$$\begin{aligned}
\langle s_i \rangle_D &= \sum_{\boldsymbol{s}} s_i(\boldsymbol{s}) \cdot p_{\hat{\boldsymbol{s}}}(\boldsymbol{s}) \\
&= \cancel{\sum_{\boldsymbol{s} \notin \hat{\boldsymbol{s}}} s_i(\boldsymbol{s}) \cdot 0} + \sum_{\boldsymbol{s} \in \hat{\boldsymbol{s}}} s_i(\boldsymbol{s}) \cdot p_{\hat{\boldsymbol{s}}}(\boldsymbol{s}) \\
&= \frac{1}{N} \sum_{\boldsymbol{s} \in \hat{\boldsymbol{s}}} s_i(\boldsymbol{s}) \cdot \#\left\{\boldsymbol{s}^{(k)} \in \hat{\boldsymbol{s}} \mid \boldsymbol{s}^{(k)} = \boldsymbol{s}\right\} \\
&= \frac{1}{N} \sum_{k=1}^{N} s_i\left(\boldsymbol{s}^{(k)}\right)
\end{aligned} \tag{20}$$

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, $\boldsymbol{s} \in \hat{\boldsymbol{s}}$ on the second and third lines is taken to mean "$\boldsymbol{s}$ occurs in the dataset", rather than being an enumeration of the rows.

Likewise for $\langle s_i s_j \rangle_D$:

$$\begin{aligned}\langle s_i s_j \rangle_D &= \sum_{\boldsymbol{s}} s_i(\boldsymbol{s}) s_j(\boldsymbol{s}) \cdot p_{\hat{\boldsymbol{s}}}(\boldsymbol{s}) \\ &= \frac{1}{N} \sum_{\boldsymbol{s} \in \hat{\boldsymbol{s}}} s_i(\boldsymbol{s}) s_j(\boldsymbol{s}) \cdot \#\left\{\boldsymbol{s}^{(k)} \in \hat{\boldsymbol{s}} \mid \boldsymbol{s}^{(k)} = \boldsymbol{s}\right\} \\ &= \frac{1}{N} \sum_{k=1}^{N} s_i\left(\boldsymbol{s}^{(k)}\right) s_j\left(\boldsymbol{s}^{(k)}\right)\end{aligned} \tag{21}$$

Q2.3. Assume that the data is stationary and that each 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 $p_{\hat{\boldsymbol{s}}}(\boldsymbol{s})$, for some microstate $\boldsymbol{s}$, as the proportion of observations in the dataset $\hat{\boldsymbol{s}}$ which were equal to $\boldsymbol{s}$. We can equivalently define $p_{\hat{\boldsymbol{s}}}(\boldsymbol{s})$ as an average over the function $\mathbb{1}\left[\boldsymbol{s}^{(k)} = \boldsymbol{s}\right]$, which is 1 if this condition holds, and 0 otherwise:

$$p_{\hat{\boldsymbol{s}}}(\boldsymbol{s}) = \frac{1}{N} \sum_{k=1}^{N} \mathbb{1}\left[\boldsymbol{s}^{(k)} = \boldsymbol{s}\right] \tag{22}$$

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

$$\lim_{N \to \infty} p_{\hat{\boldsymbol{s}}}(\boldsymbol{s}) = p(\boldsymbol{s}) \tag{23}$$

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

$$\lim_{N \to \infty} \langle s_i \rangle_D = \lim_{N \to \infty} \left[\sum_{\boldsymbol{s} \in \hat{\boldsymbol{s}}} s_i(\boldsymbol{s}) \cdot p_{\hat{\boldsymbol{s}}}(\boldsymbol{s}) \right] = \sum_{\boldsymbol{s}} s_i(\boldsymbol{s}) \cdot p(\boldsymbol{s}) = \langle s_i \rangle \tag{24}$$

And analogously for the local pairwise correlation:

$$\lim_{N \to \infty} \langle s_i s_j \rangle_D = \lim_{N \to \infty} \left[\sum_{\boldsymbol{s} \in \hat{\boldsymbol{s}}} s_i(\boldsymbol{s}) s_j(\boldsymbol{s}) \cdot p_{\hat{\boldsymbol{s}}}(\boldsymbol{s}) \right] = \sum_{\boldsymbol{s}} s_i(\boldsymbol{s}) s_j(\boldsymbol{s}) \cdot p(\boldsymbol{s}) = \langle s_i s_j \rangle \tag{25}$$

2.3 Maximum Entropy models

In many papers, the authors refer to the generalised Ising model defined in Equation 9 as a *maximum entropy model*. In this section we will show that the probability distribution defined in Equation 9 is indeed the (most general) probability distribution that maximises the Shannon entropy $S\left[p_{\boldsymbol{g}}(\boldsymbol{s})\right]$ given a set of constraints (which we will specify).

Q3.1. 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:

$$S[p(\boldsymbol{s})] = -\sum_{\boldsymbol{s}} p(\boldsymbol{s}) \log p(\boldsymbol{s}) \tag{26}$$

where the summation is over all possible microstates $\boldsymbol{s}$.

The Ising model in Equation 9 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

$$\sum_{\boldsymbol{s}} p(\boldsymbol{s}) = 1 \quad \text{and} \quad \sum_{\boldsymbol{s}} p(\boldsymbol{s}) s_i(\boldsymbol{s}) = \langle s_i \rangle_D \quad \text{and} \quad \sum_{\boldsymbol{s}} p(\boldsymbol{s}) s_i(\boldsymbol{s}) s_j(\boldsymbol{s}) = \langle s_i s_j \rangle_D \tag{27}$$

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).

Q3.2. How many constraints are there in total?

The constraints are as follows:

- **Normalisation:** 1 constraint,
- **Average local magnetisation:** n (one per spin)
- **Average local correlation:** $\frac{n(n-1)}{2}$ (one for each pair of distinct spins)

Thus the total number of constraints is

$$1 + n + \frac{n(n-1)}{2} = \frac{n(n+1)}{2} + 1 \tag{28}$$

i.e., the length of the parameter vector $\boldsymbol{g}$, with an additional normalisation constraint to impose interdependence on the elements of $\boldsymbol{g}$.

Note that *maximisation* could be considered a constraint (though this is is a somewhat non-standard view in the context of constrained optimisation), in which case the total number of constraints increases by 1.

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

$$\begin{aligned} U[p(\boldsymbol{s})] = S[p(\boldsymbol{s})] &+ \lambda_0 \left(\sum_{s} p(\boldsymbol{s}) - 1 \right) + \sum_{i=1}^{n} \alpha_i \left(\sum_{\boldsymbol{s}} p(\boldsymbol{s}) s_i(\boldsymbol{s}) - \langle s_i \rangle_D \right) \\ &+ \sum_{\text{pair}(i,j)}^{n} \eta_{ij} \left(\sum_{\boldsymbol{s}} p(\boldsymbol{s}) s_i(\boldsymbol{s}) s_j(\boldsymbol{s}) - \langle s_i s_j \rangle_D \right) \end{aligned} \tag{29}$$

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 $p(\boldsymbol{s})$.

Q3.3. 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:

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

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 $p_{\boldsymbol{s}}$ in the vector $\boldsymbol{p}$, in each of the sums over $\boldsymbol{s}'$ in U, all terms will be annhialated by the derivative, with the exception of $\boldsymbol{s}' = \boldsymbol{s}$.

First, examining the Shannon entropy:

$$\begin{aligned} \frac{\partial}{\partial p_{\boldsymbol{s}}} S[p(\boldsymbol{s})] &= \frac{\partial}{\partial p_{\boldsymbol{s}}} \left[-\sum_{\boldsymbol{s}'} p(\boldsymbol{s}') \log p(\boldsymbol{s}') \right] \\ &= -\log(p_{\boldsymbol{s}}) - p_{\boldsymbol{s}} \cdot \frac{1}{p_{\boldsymbol{s}}} \\ &= -\log(p_{\boldsymbol{s}}) - 1 \end{aligned} \tag{31}$$

Next, the normalisation constraint:

$$\frac{\partial}{\partial p_{\boldsymbol{s}}}\lambda_0\left(\sum_{\boldsymbol{s}'}p(\boldsymbol{s}')-1\right)=\frac{\partial}{\partial p_{\boldsymbol{s}}}\lambda_0 p_{\boldsymbol{s}}=\lambda_0 \tag{32}$$

The local average magnetisation constraint:

$$\begin{aligned}\frac{\partial}{\partial p_{\boldsymbol{s}}}\sum_{i=1}^{n}\alpha_i\left(\sum_{\boldsymbol{s}'}p(\boldsymbol{s}')s_i(\boldsymbol{s}')-\langle s_i\rangle_D\right)&=\sum_{i=1}^{n}\alpha_i\left(\frac{\partial}{\partial p_{\boldsymbol{s}}}p_{\boldsymbol{s}}s_i(\boldsymbol{s})\right)\\&=\sum_{i=1}^{n}\alpha_i s_i(\boldsymbol{s})\end{aligned} \tag{33}$$

And lastly the average local correlation constraint:

$$\begin{aligned}\frac{\partial}{\partial p_{\boldsymbol{s}}}\sum_{\text{pair}(i,j)}\eta_{ij}\left(\sum_{\boldsymbol{s}'}p(\boldsymbol{s}')s_i(\boldsymbol{s}')s_j\boldsymbol{s}'-\langle s_i s_j\rangle_D\right)&=\sum_{\text{pair}(i,j)}\eta_{ij}\left(\frac{\partial}{\partial p_{\boldsymbol{s}}}p_{\boldsymbol{s}}s_i(\boldsymbol{s})s_j(\boldsymbol{s})\right)\\&=\sum_{\text{pair}(i,j)}\eta_{ij}s_i(\boldsymbol{s})s_j(\boldsymbol{s})\end{aligned} \tag{34}$$

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

$$\frac{\partial U[\boldsymbol{p}]}{\partial p_{\boldsymbol{s}}}=-\log(p_{\boldsymbol{s}})-1+\lambda_0+\sum_{i=1}^{n}\alpha_i s_i(\boldsymbol{s})+\sum_{\text{pair}(i,j)}\eta_{ij}s_i(\boldsymbol{s})s_j(\boldsymbol{s}) \tag{35}$$

Q3.4. Can you show that the most general expression of $p_{\boldsymbol{s}}$ with maximal entropy that satisfying the constraints in Equation 27 is Equation 9? 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,

$$\frac{\partial U}{\partial p_s}=\frac{\partial U}{\partial \lambda_0}=\frac{\partial U}{\partial \alpha_i}=\frac{\partial U}{\partial \eta_{ij}}=0 \tag{36}$$

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_s is zero for a particular $\boldsymbol{s}$ when

$$\begin{aligned}0&=-\log(p_{\boldsymbol{s}})-1+\lambda_0+\sum_{i=1}^{n}\alpha_i s_i(\boldsymbol{s})+\sum_{\text{pair}(i,j)}\eta_{ij}s_i(\boldsymbol{s})s_j(\boldsymbol{s})\\p_{\boldsymbol{s}}&=\exp\left(-(1-\lambda_0)+\sum_{i=1}^{n}\alpha_i s_i(\boldsymbol{s})+\sum_{\text{pair}(i,j)}\eta_{ij}s_i(\boldsymbol{s})s_j(\boldsymbol{s})\right)\\&=\frac{1}{\exp(1-\lambda_0)}\exp\left(\sum_{i=1}^{n}\alpha_i s_i(\boldsymbol{s})+\sum_{\text{pair}(i,j)}\eta_{ij}s_i(\boldsymbol{s})s_j(\boldsymbol{s})\right)\end{aligned} \tag{37}$$

Equating Equation 9 with Equation 37, we obtain the following relationships:

$$\begin{aligned}Z&=\exp(1-\lambda_0)\\h_i&=\alpha_i&&\text{for each spin } i\\J_{ij}&=\eta_{ij}&&\text{for each pair of spins } i,j\end{aligned} \tag{38}$$

2.4 Statistical inference: model with no couplings

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

$$p_{\boldsymbol{g}}(\boldsymbol{s}) = \frac{1}{Z(\boldsymbol{g})} \exp\left(\sum_{i=1}^{n} h_i s_i\right) \tag{39}$$

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

Q4.1. 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:

$$\begin{aligned} Z(\boldsymbol{g}) &= \sum_{\boldsymbol{s}} \exp\left(\sum_{i=1}^{n} h_i s_i\right) \\ &= \sum_{s_1 \in \pm 1} \cdots \sum_{s_{n-1} \in \pm 1} \exp\left(\sum_{i=1}^{n-1} h_i s_i\right) (\exp(h_n) + \exp(-h_n)) \\ &= \prod_{i=1}^{n} \exp(h_i) + \exp(-h_i) \\ &= \prod_{i=1}^{n} 2\cosh(h_i) \end{aligned} \tag{40}$$

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

$$\begin{aligned} p_{\boldsymbol{g}}(\boldsymbol{s}) &= \frac{1}{Z(\boldsymbol{g})} \exp\left(\sum_{i=1}^{n} h_i s_i\right) \\ &= \frac{1}{Z(\boldsymbol{g})} \prod_{i=1}^{n} \exp(h_i s_i) \\ &= \prod_{i=1}^{n} \frac{\exp(h_i s_i)}{2\cosh(h_i)} \end{aligned} \tag{41}$$

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:

$$p_{\boldsymbol{h_i}}(s_i) = \frac{\exp(h_i s_i)}{2\cosh(h_i)} \tag{42}$$

Q4.2. 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:

$$h_i = \tanh^{-1}(\langle s_i \rangle_D), \tag{43}$$

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 $\langle s_i \rangle$

$$\langle s_i \rangle = \sum_{\boldsymbol{s}} s_i(\boldsymbol{s}) \cdot p_{\boldsymbol{g}}(\boldsymbol{s}) = \langle s_i \rangle_D \tag{44}$$

From the previous exercise, we have that $p_{\boldsymbol{g}}(\boldsymbol{s}) = \prod_{j=1}^{n} p_{\boldsymbol{h}_j}(s_j)$. Taking $\boldsymbol{s}_{-i} = (s_1, ..., s_{i-1}, s_{i+1}, ..., s_n)$ to denote the vector of spin variables excluding s_i, we use this result to rewrite Equation 44 in terms of $p_{\boldsymbol{h}_i}$:

$$\begin{aligned}
\langle s_i \rangle_D &= \sum_{\boldsymbol{s}} s_i(\boldsymbol{s}) \cdot p_{\boldsymbol{h}_i}(s_i(\boldsymbol{s})) \cdot \prod_{j \neq i} p_{\boldsymbol{h}_j}\big(s_j(\boldsymbol{s})\big) \\
&= \sum_{s \in \pm 1} s p_{\boldsymbol{h}_i}(s) \cdot \overbrace{\sum_{\boldsymbol{s}_{-i}} \prod_{j \neq i} p_{\boldsymbol{h}_j}\big(s_j(\boldsymbol{s}_{-i})\big)}^{=1} \\
&= p_{\boldsymbol{h}_i}(1) - p_{\boldsymbol{h}_i}(-1) \\
&= \frac{\exp(h_i) - \exp(-h_i)}{2\cosh(h_i)} \\
&= \frac{\sinh(h_i)}{\cosh(h_i)} \\
&= \tanh(h_i)
\end{aligned} \tag{45}$$

In the second line of reasoning we group the original summation terms by their s_i component, conditional on observing any other combination of states in $\boldsymbol{s}_{-i}$. Since each $p_{\boldsymbol{h}_j}$ is an independent probability distribution, it follows that the sum over the joint probabilities of $\boldsymbol{s}_{-i}$ is normalised, hence allowing the simplification in the third line.

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

Q4.3. 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 Q1.4, in which we interpreted the sign of h_i as reflecting 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 have made this assumption explicit by taking $J_{ij} = 0$.

2.5 Statistical inference: maximising the log-likelihood function

Introducing the likelihood function. Looking more closely at Equation 9, 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 Equation 9 changes continuously as one continuously varies the parameters in $\boldsymbol{g}$. We say that Equation 9 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*:

$$\mathcal{L}(\boldsymbol{g}) = \log P_{\boldsymbol{g}}(\boldsymbol{s}) \tag{46}$$

where $P_{\boldsymbol{g}}(\boldsymbol{s})$ is the probability that the model $p_{\boldsymbol{g}}(\boldsymbol{s})$ produces the dataset $\hat{\boldsymbol{s}} = \big(\boldsymbol{s}^{(1)}, ..., \boldsymbol{s}^{(N)}\big)$. 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 $\boldsymbol{g}$.

Q5.1. We assume that, in the dataset $\hat{\boldsymbol{s}}$, all datapoints are independently sampled from an underlying distribution $p_{\boldsymbol{g}}(\boldsymbol{s})$. Can you show that the log-likelihood function can be re-written as:

$$\mathcal{L}(\boldsymbol{g}) = N \sum_{\boldsymbol{s}} p_D(\boldsymbol{s}) \log p_{\boldsymbol{g}}(\boldsymbol{s}) \tag{47}$$

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 $P_{\boldsymbol{g}}(\hat{\boldsymbol{s}}) = \prod_{k=1}^{N} p_{\boldsymbol{g}}\big(\hat{\boldsymbol{s}}^{(k)}\big)$. Substituting this in Equation 46, we have:

$$\begin{aligned}
\mathcal{L}(\boldsymbol{g}) &= \log\left(\prod_{k=1}^{N} p_{\boldsymbol{g}}\big(\hat{\boldsymbol{s}}^{(k)}\big)\right) \\
&= \sum_{k=1}^{N} \log p_{\boldsymbol{g}}\big(\hat{\boldsymbol{s}}^{(k)}\big) \\
&= \sum_{\boldsymbol{s}} K(\boldsymbol{s}) \log p_{\boldsymbol{g}}(\boldsymbol{s}) \\
&= \sum_{\boldsymbol{s}} N \cdot \frac{K(\boldsymbol{s})}{N} \log p_{\boldsymbol{g}}(\boldsymbol{s}) \\
&= N \sum_{\boldsymbol{s}} p_D(\boldsymbol{s}) \log p_{\boldsymbol{g}}(\boldsymbol{s}) \qquad (48)
\end{aligned}$$

as desired, where the summation is over uniques microstates in the dataset (or equivalently over all possible microstates).

Ising model. We now take the model distribution $p_{\boldsymbol{g}}(\boldsymbol{s})$ to be given by the Ising model in Equation 9.

Q5.2. 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 Equation 9, and let i correspond to the spin s_i. Then $\mathcal{L}(\boldsymbol{g})$ is:

$$\begin{aligned}
\mathcal{L}(\boldsymbol{g}) &= N \sum_{\boldsymbol{s}} p_D(\boldsymbol{s}) \log\left[\frac{1}{Z(\boldsymbol{g})} \exp\left(\sum_{j=1}^{n} h_j s_j + \sum_{\text{pair}(j,k)} J_{jk} s_j s_k\right)\right] \\
&= N \sum_{\boldsymbol{s}} p_D(\boldsymbol{s}) \left[\sum_{j=1}^{n} h_j s_j + \sum_{\text{pair}(j,k)} J_{jk} s_j s_k - \log Z(\boldsymbol{g})\right] \qquad (49)
\end{aligned}$$

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

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial h_i} &= N \sum_{\boldsymbol{s}} p_D(\boldsymbol{s}) \left[\frac{\partial}{\partial h_i} \sum_{j=1}^{n} h_j s_j + \frac{\partial}{\partial h_i} \sum_{\text{pair}(j,k)} J_{jk} s_j s_k - \frac{\partial}{\partial h_i} \log Z(\boldsymbol{g})\right] \\
&= N \sum_{\boldsymbol{s}} p_D(\boldsymbol{s}) \cdot \left(s_i - \frac{\partial}{\partial h_i} \log Z(\boldsymbol{g})\right) \qquad (50)
\end{aligned}$$

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

$$\begin{aligned}
\frac{\partial \log Z(\boldsymbol{g})}{\partial h_i} &= \frac{1}{Z(\boldsymbol{g})} \cdot \frac{\partial}{\partial h_i} \sum_{\boldsymbol{s}'} \exp\left(\sum_{j=1}^{n} h_j s_j + \sum_{\text{pair}(j,k)} J_{jk} s_j s_k\right) \\
&= \frac{1}{Z(\boldsymbol{g})} \cdot \sum_{\boldsymbol{s}'} \frac{\partial}{\partial h_i} \exp\left(\sum_{j=1}^{n} h_j s_j + \sum_{\text{pair}(j,k)} J_{jk} s_j s_k\right) \\
&= \frac{1}{Z(\boldsymbol{g})} \cdot \sum_{\boldsymbol{s}'} \exp\left(\sum_{j=1}^{n} h_j s_j + \sum_{\text{pair}(j,k)} J_{jk} s_j s_k\right) \cdot s_i \\
&= \sum_{\boldsymbol{s}'} p_{\boldsymbol{g}}(\boldsymbol{s}') s_i \\
&= \langle s_i \rangle \qquad (51)
\end{aligned}$$

$\mathcal{L}(\boldsymbol{g})$ attains its maximum value with respect to h_i when $\frac{\partial \mathcal{L}}{\partial h_i} = 0$. Substituting Equation 51 into Equation 50 and solving for the maximum, we find:

$$\begin{aligned} 0 &= N \sum_{\boldsymbol{s}} p_D(\boldsymbol{s}) \cdot (s_i - \langle s_i \rangle) \\ \implies \quad N \sum_{\boldsymbol{s}} p_D(\boldsymbol{s}) \langle s_i \rangle &= N \sum_{\boldsymbol{s}} p_D(\boldsymbol{s}) s_i \\ \langle s_i \rangle \cdot \sum_{\boldsymbol{s}} p_D(\boldsymbol{s}) &= \langle s_i \rangle_D \\ \langle s_i \rangle &= \langle s_i \rangle_D \end{aligned} \tag{52}$$

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}:

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial J_{ij}} &= N \sum_{\boldsymbol{s}} p_D(\boldsymbol{s}) \left[\frac{\partial}{\partial J_{ij}} \sum_{k=1}^{n} h_k s_k + \frac{\partial}{\partial J_{ij}} \sum_{\text{pair}(k,\ell)} J_{k\ell} s_k s_\ell - \frac{\partial}{\partial J_{ij}} \log Z(\boldsymbol{g}) \right] \\ &= N \sum_{\boldsymbol{s}} p_D(\boldsymbol{s}) \cdot \left(s_i - \frac{\partial}{\partial h_i} \log Z(\boldsymbol{g}) \right) \\ &= N \sum_{\boldsymbol{s}} p_D(\boldsymbol{s}) \cdot \left(s_i - \sum_{\boldsymbol{s}'} p_{\boldsymbol{g}}(\boldsymbol{s}') s_i s_j \right) \\ &= N \sum_{\boldsymbol{s}} p_D(\boldsymbol{s}) \cdot \left(s_i - \langle s_i s_j \rangle \right) \end{aligned} \tag{53}$$

As before, evaluating Equation 53 at $\frac{\partial \mathcal{L}}{\partial J_{ij}} = 0$ gives:

$$\begin{aligned} N \sum_{\boldsymbol{s}} p_D(\boldsymbol{s}) \langle s_i s_j \rangle &= N \sum_{\boldsymbol{s}} p_D(\boldsymbol{s}) s_i s_j \\ \langle s_i s_j \rangle &= \langle s_i s_j \rangle_D \end{aligned} \tag{54}$$

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.

Q1. 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?

The number of spins corresponds to the number of judges ($n = 9$), for which there are $2^9 = 512$ possible combinations of votes. The dataset contains $895 > 512$ rows; however, only $N_{\text{max}} = 128$ of these are unique.

Q2. (Bonus question) Numerical solution: For the dataset provided, find numerically the value of the parameters $\boldsymbol{g}$ of the fully-connected pairwise model Equation 9 that maximises the log-likelihood function $\log \mathcal{L}(\boldsymbol{g})$. What are the main computational limitations of your algorithm? How can you improve it?

We solve for the parameters $\boldsymbol{g}$ by minimising the negative log-likelihood function (Equation 47) using the SciPy BFGS solver (line 597 of `combined.py` or line 386 of `supreme_court.py`). Note that this is equivalent to *maximising* the log-likelihood function. The negative log-likelihood across optimisation iterations is shown in Figure 4.

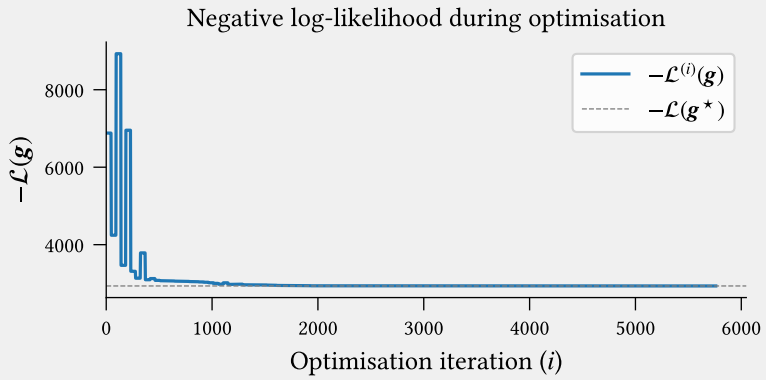


Figure 4: Negative log-likelihood during fitting of pairwise Ising model to the Supreme Court data, using the SciPy BFGS solver subroutine. The grey dashed line indicates the negative log-likelihood of the parameters $\boldsymbol{g}^\star$ provided on Canvas.

While we obtain similar fit parameters to those provided on Canvas, the optimisation process is computationally expensive – even for only nine judges – in particular because the partition function must be re-evaluated at each step. Computing $Z(g)$ requires a summation over 2^n distinct microstates, with $O(n^2)$ operations for each term.

One approach to alleviating this computational limitation is to approximate the distribution $p_{\boldsymbol{g}}(\boldsymbol{s})$ using Metropolis sampling, such that we don't need to calculate the partition function. This dataset contains a small number of states which occur with relatively high probability – the two most common microstates absorb 45% of the probability mass. As such, there is reason to expect a reasonably accurate estimate for $p_{\boldsymbol{g}}(\boldsymbol{s})$ from a small number of samples (compared to the number of microstates).

Q3. We would like to reproduce **Figure 13** of the attached paper. Can you plot the $\langle s_i \rangle_D$ as a function of i? Can you re-order the label i so that the values of $\langle s_i \rangle_D$ are ordered from the smallest (negative value) to the largest (positive value), as in Fig. 13.A top? Keeping the new ordering, can you plot a heatmap of the matrix of $\langle s_i s_j \rangle_D$ (see Fig. 13.A bottom)? What can we say about the judges with negative $\langle s_i \rangle_D$? With positive $\langle s_i \rangle_D$? Can you comment on these plots?

We calculate $\langle s_i \rangle_D$ and $\langle s_i s_j \rangle_D$ using the methods derived in Equation 20 and Equation 21. The results, in increasing order of $\langle s_i \rangle_D$, are displayed in Figure 5 and Figure 6 respectively.

In Figure 5 we observe a 5-4 split between judges who vote mostly conservative, and those who mostly vote liberal. We also note that $|\langle s_i \rangle_D| < 0.5$ for all judges, indicating that even the most conservative/liberal judges vote against their ideological leaning in at least 25% of cases.

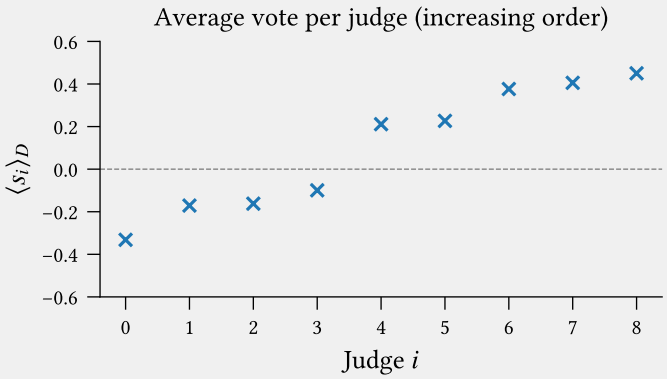


Figure 5: Average vote per judge in the supplied dataset, with x-axis ordered according to increasing tendency to vote conservative. Votes in the dataset are assigned -1 if considered **liberal**, and $+1$ if considered **conservative**.

Figure 6 also exhibits several features of interest. The liberal and conservative blocks are clearly visible; however, no two judges are completely aligned. Within the liberal block, the first judge (who votes liberal most often) is distinct from the other three. A similar separation appears to exist in the conservative block as well, between the first three and final two judges – this is less evident in the context of the conservative block alone, yet we see the final two judges show extremely low correlation with the most liberal judge.

A particularly counter-intuitive feature of Figure 6 is that all observed values are positive. We observed in Figure 5 that some judges predominantly vote liberally, and others conservatively. If we suppose that the judges' votes are independent of one another, we would expect to see a negative correlation between liberal-conservative pairs of judges. Since we observe *no negative correlations*, this implies that the votes are not independent, and that interaction between judges may play a significant role in any particular judge's vote – irrespective of their political leaning.

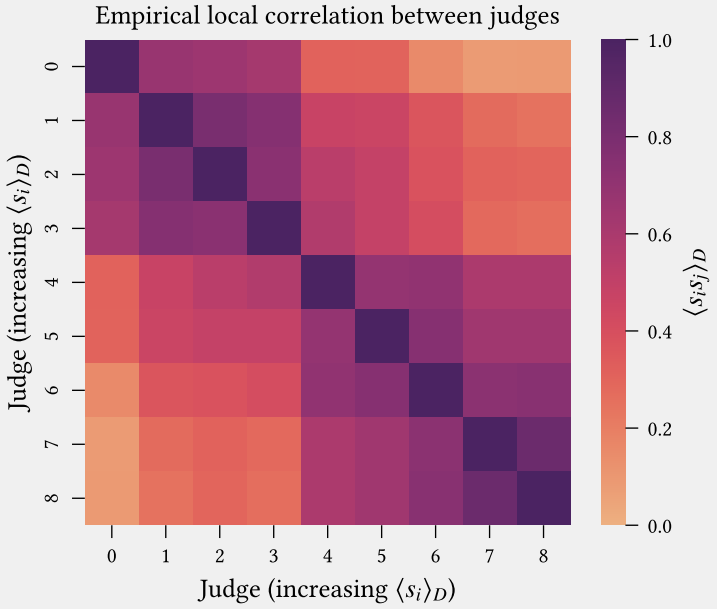


Figure 6: Pairwise correlations between judges' votes. Axes ordered by increasing average tendency to vote conservative.

Critically, we note that the pairwise correlations in Figure 6 are not sufficient for extended discussion on the interactions between judges, as the values include the heterogeneous influence of individual judges' voting patterns.

Q4. Keeping the new ordering of the labels i, can you plot a heatmap of the fitted vector of h_i's and a heatmap of the fitted matrix of J_{ij}'s, as in Fig. 13.B? You can use the fitted values of h_i and J_{ij} provided in Canvas. Can you comment on these plots?
Note that the values of h_i and J_{ij} that are provided in Canvas are following the same order of the variables s_i than in the USSC datafile (i.e., the original ordering of the judges).

The fit h_i parameters are displayed in Figure 7, and the interaction terms J_{ij} in Figure 8. In both cases, the judges are sorted in their axes in accordance with increasing order of $\langle s_i \rangle_D$, the tendency for a judge to vote conservatively. We observe substantial differences between each figure and its counterpart in the previous question.

The h_i values displayed in Figure 7 reflect the political leaning of each judge, in absence of interactions. We observe several surprising inversions with respect to $\langle s_i \rangle_D$. For instance, Figure 5 displayed a tendency for the third judge to vote (on average) more liberally, yet $h_3 > 0$ indicates a conservative leaning. On the other hand, h_5 and h_7 are both close to 0, yet these judges' votes displayed non-neglibible conservative tendencies.

To understand this counter-intuitive result, we must consider it in the context of the model interaction terms. Indeed, in **Q1.4** we interpreted h_i as the political leaning of the i'th judge, *in absence of interactions*.

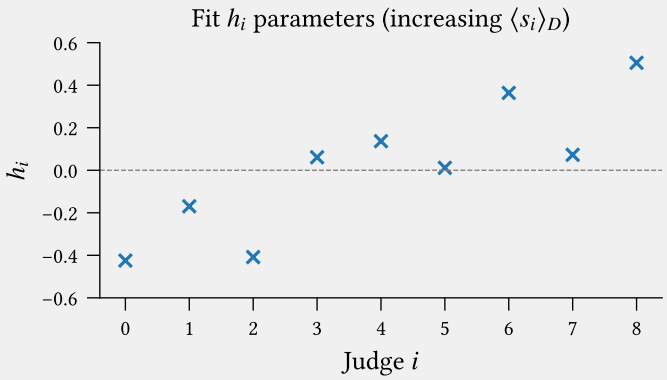


Figure 7: Fit h_i parameter values per judge in the supplied dataset, with x-axis sorted according to increasing tendency to vote conservative. Votes in the dataset are assigned -1 if considered **liberal** and $+1$ if considered **conservative**.

The fit interaction terms are displayed in Figure 8. The values J_{ij} indicate the tendency for a pair of judges to vote in-line or against one another, regardless of their political leaning. A value close to zero implies that any similarity/dissimilarity in the votes of a pair of judges is well-explained by their political leanings (or interactions with other judges). A positive value suggests that two judges may vote the same way often, even when this contradicts ones' political leaning. Note that the diagonal terms are 0 on account of the model disregarding self-interactions (i.e., such parameters do not exist in the model).

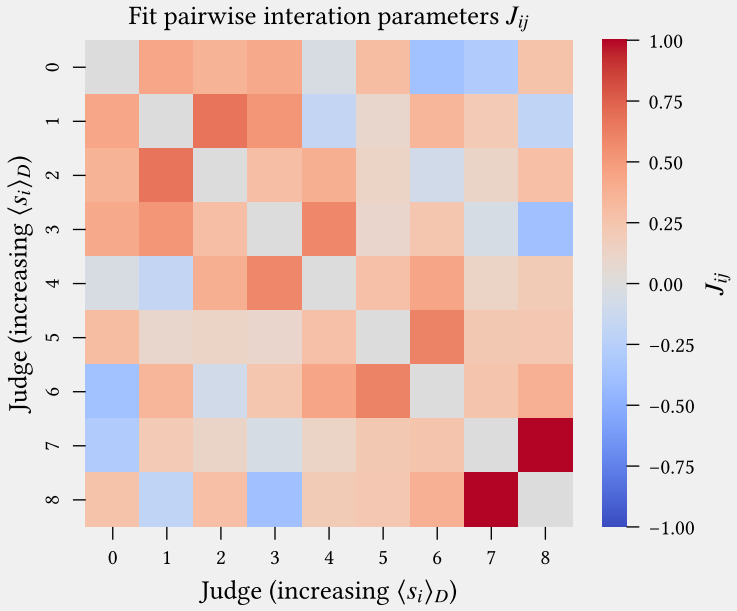


Figure 8: Interaction terms J_{ij} between judges in a pairwise Ising model fit to the supplied dataset. Axes are ordered according to increasing tendency for a judge to vote conservatively.

We observe qualitatively different behaviour with respect to the local correlations in Figure 6. Firstly, the heatmap now includes negative terms, though these typically have lower magnitude that the positive interactions. The liberal and conservative "blocks" identified earlier each display positive pairwise interaction values, with negative interactions restricted to cross-block pairs of judges. Interestingly, we also observe some positive interactions between the blocks, which vary between judges. For instance, the 0'th judge has a positive interaction value with judges 5 and 8, while the 1'st judge displays a negative interaction with the 8′th judge, but positive interactions with the 6′th and 7′th. Overall we observe considerably more heterogenerous behaviour in Figure 8 than in Figure 6.

Returning to our examination of judges 3, 5, and 7, Figure 8 offers some insight into their unexpected inferred political leanings. Consider the 3rd judge. In the data, they display a tendency to vote liberally, yet $h_3 > 0$ indicates that this tendency does not arise due to their political leaning. For the model to accurately reflect the data (we will see in later discussion that it does), this behaviour must then arise from the 3rd judge's interactions with other judges.

Indeed, we see from Figure 8 that they exhibit positive interactions with judges 0, 1 and 2, and negative interaction values with judges 7 and 8 (who we earlier predicted to be the most conservative). Intuitively, this implies that the 3rd judge tends to vote similarly to liberal judges, but differently to the more conservative judges. Thus their liberal voting tendencies are explained not by their political leaning, but by their tendency to vote "like a liberal judge".

The 5′th judge actually exhibits positive interaction values with all judges, yet their strongest interaction is with the 6′th judge, who in turn consistently votes opposite to the 0′th judge (the most liberal), explaining the tendency for the 5′th judge to vote conservative. Likewise, the 7′th judge – who we had predicted as one of the more conservative – displays only a small conservative leaning. This is explained by their strong positive interactions with the 8′th judge, and tendency to vote oppositely to the 0′th judge.

Q5. Scatter plot 1, "Cross-validation": For all the states observed in the data, can you plot the empirical probability of the state, $p_D(\boldsymbol{s})$ against the model probability $p_{\boldsymbol{g}}(\boldsymbol{s})$ with the fitted parameters (h_i and J_{ij})? What can you say about this plot?

Figure 9 compares the probability of observed microstates under the fit Ising model with that observed in the dataset. We see general agreement between the two. In particular, the most probable states in the dataset have similar probability under the fit model.

The discrepancies indicate potential higher-order interactions which are not captured by the pairwise Ising model. Since this model is constrained to reproduce $\langle s_i \rangle_D$ and $\langle s_i s_j \rangle_D$ for each s_i, s_j, these discrepancies cannot be explained by the political leaning of particular judges or pairwise interactions alone. However, such discrepancies appear relatively minor, suggesting that the pairwise model is a reasonably good explanation for the data.

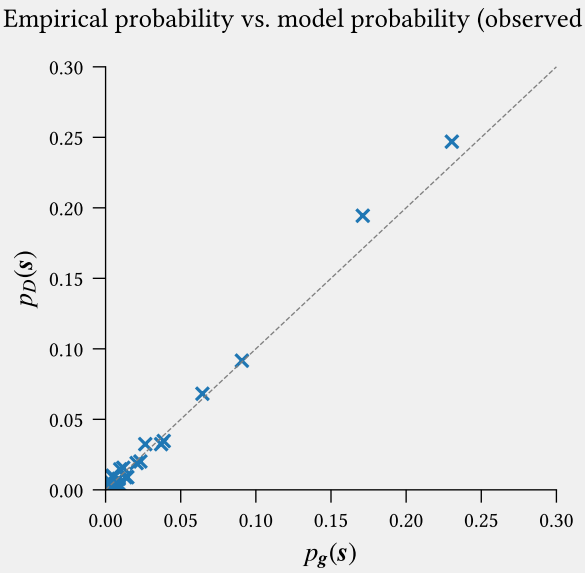


Figure 9: Cross-validation check: comparison between microstate probability under the fit pairwise Ising model, and the observed data. The $y = x$ line is indicated by a grey dashed line.

Q6. Scatter plot 2: Checking the fit: For each spin s_i, can you plot the value of $\langle s_i \rangle_D$ in the data against the value $\langle s_i \rangle$ in the fitted model? For each pair of spins s_i, s_j, can you plot the value of $\langle s_i s_j \rangle_D$ in the data against the value of $\langle s_i s_j \rangle$ in the fitted model? What can you say about these plots?

Figure 10 and Figure 11 compare the model average magnetisation and pairwise correlation with those values calculated from the observed data. We observe strong agreement between the two. Following our derivation in **Q5.2**, this implies that the model parameters are well-fit to the data, and in particular that the log-likelihood is maximal.

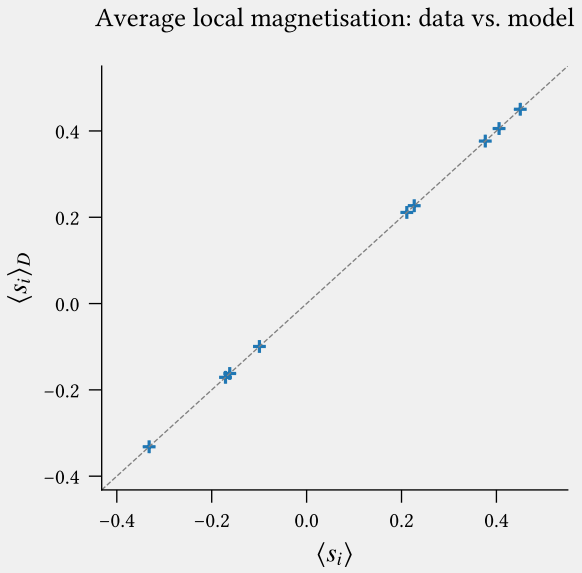


Figure 10: Comparison of average local magnetisation under the fit model and as calculated from the observed data. The $y = x$ line is indicated in grey. We observe that the model accurately reproduces the data value.

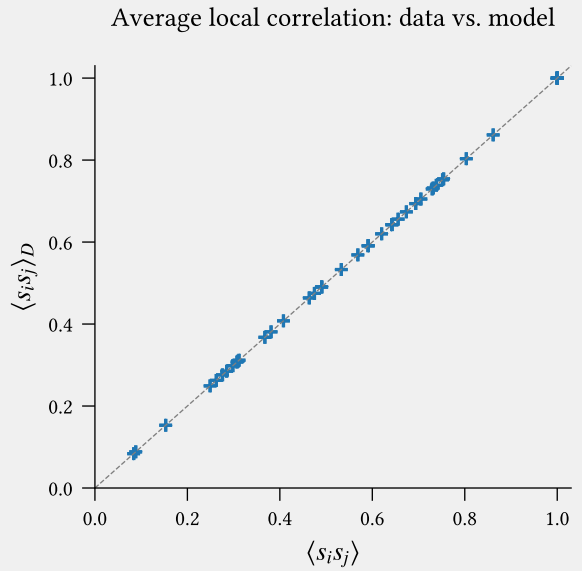


Figure 11: Comparison of average local pairwise correlation under the fit model and as calculated from the observed data. The $y = x$ line is indicated in grey. We observe that the model accurately reproduces the data value.

One of the questions addressed in the paper is: can the pairwise model reproduce higher-order patterns of the data better than a model of independent judges? To answer this question, the authors introduce the probability $P(k)$ that there are k -conservative votes as answer to a case (i.e., the probability that a datapoint contains k conservative votes).

Q7. Consider a model with no coupling, in which judges vote independently from each other. Each judge s_i has a probability $p_i(+1)$ to vote conservative. In that case, what is the probability $P_I(k)$ that k judges vote conservative? Note: we added the label "I" to $P(k)$ to specify that it is the probability distribution obtained for an independent model.
In the dataset, how many judges have votes that are more conservative on average? At which value of k do you then expect the maximum of $P_I(k)$ to be? Can you obtain the values of $p_i(+1)$ from the data and plot the values of $P_I(k)$ as a function of k for the independent model?

From Figure 5 there are 5 judges whose votes are, on average, more conservative. Since $P_I(k)$ assumes that judges vote independently, we then expect a maximum at $k = 5$. This makes intuitive sense, since $k = 6$ requires some judges who

vote liberally to vote conservatively, and $k = 4$ would require an (on-average) conservative judge to vote liberally. These are both less likely than the case where each judge votes in-line with their typical behaviour. Note that the peak at $k = 5$ does include mass from scenarios where typically-liberal judges vote conservative, and vice versa.

It is straightforward to obtain $p_i(+1)$ from the data as the proportion of rows where each judge votes conservative. It is less straightforward to compute $P_I(k)$ given each $p_i(+1)$, as k conservative votes could come from any one of $\binom{9}{k}$ combinations of judges.

One approach to simplifying the calculation of $P_I(k)$ is to redefine it as a recurrence relation which can be computed straightforwardly using recursion or dynamic programming. We define the probability $q(x, i)$ as the probability of observing x conservative votes among the final i judges (under an arbitrary ordering of judges). Formally, we define q by:

$$q(x,i) = \begin{cases} 0 & \text{if } x > i \\ \prod_{j=(n-i)+1}^{n} p_j(-1) & \text{if } x = 0 \\ \prod_{j=(n-i)+1}^{n} p_j(+1) & \text{if } x = i \\ p_{(n-i)+1}(+1) \cdot q(x-1, i-1) + p_{(n-i)+1}(-1) \cdot q(x, i-1) & \text{if } x < i \end{cases} \tag{55}$$

In the first case ($x > i$), we require more conservative votes than remaining judges, which naturally has probability zero. In the second ($x = 0$), any remaining votes must be liberal, as we have achieved the required number of conservative votes. Likewise, in the third case there are exactly as many judges remaining as required conservative votes, so all must vote conservative.

Finally, in the last case, we consider two scenarios which are recursively defined: where the current judge votes conservative, requiring one fewer conservative votes among the remaining judges, or where the current judge votes liberal, and we still require x conservative votes.

To avoid possible floating-point innacuracies when multiplying small probabilities, the products are implemented as the exponentiated sum of log-probabilities.

The distribution $P_I(k) = q(k, n)$ is shown below in Figure 12.

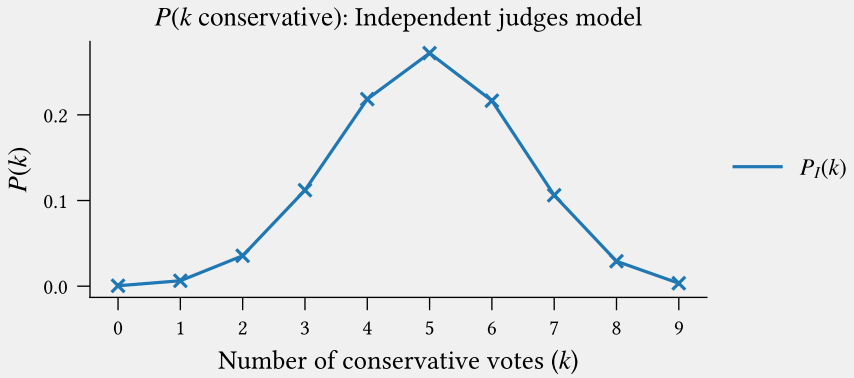


Figure 12: Distribution of the number of conservative votes k, calculated with respect to the (assumed independent) probabilities of each individual judge voting conservative.

Q8. Let us call $P_D(k)$ the probability distribution $P(k)$ obtained directly from the data. How can you compute the values of $P_D(k)$ from the data for $k = 1$? for $k = 2$? for any k? Can you plot $P_D(k)$ as a function of k and compare the curve to the one obtained previously for the independent model? Where is the maximum of $P_D(k)$? Is the independent model a good model for the US Supreme Court data?

Observe that in a model with n judges, if $k \leq n$ judges vote conservative, then with the $\{-1, +1\}$ valuation on votes, the sum of the votes gives a net vote of

$$1 \cdot k + (-1) \cdot (n - k) = 2k - n \tag{56}$$

It is then straightforward to calculate $P_D(k)$ for any k as the proportion of rows in D whose values sum to $2k - n$. The resulting distribution is shown with comparison to $P_I(k)$ in Figure 13.

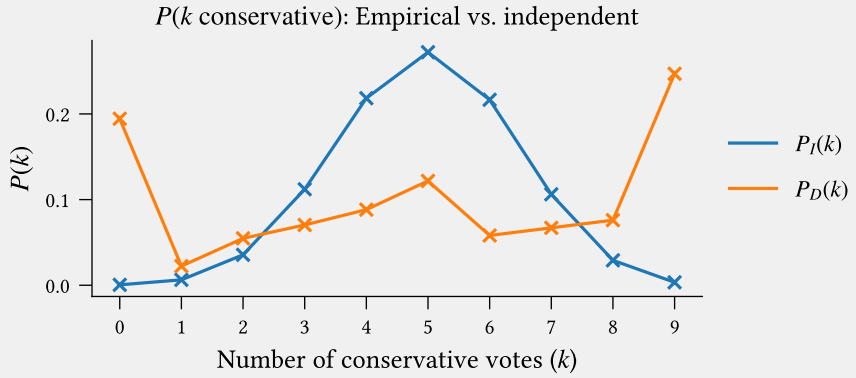


Figure 13: Comparison between distribution of microstates in the data (orange) with the distribution predicted under the assumption that judges' votes are independent of one another (blue).

We observe considerable differences between the two distributions. While $P_D(k)$ still has a local maximum at $k = 5$, this state (five conservative votes) is significantly less likely than predicted by the independent model. Additionally, $P_D(k)$ exhibits two other local maxima at $k = 0$ and $k = 9$, with the latter also being the global maximum.

A plausible (but incorrect) interpretation of these maxima is that they correspond to issues which are "too liberal" or "too conservative" for all judges. However, given the substantial probability mass at these peaks, if this were the case we would not observe such low probability for these cases in $P_I(k)$.

More likely, it is the case that these peaks correspond to issues where judges' votes cannot be explained entirely by their political leaning, i.e., that interactions between judges has resulted in agreement, sufficient to cause liberal judges to vote conservative, or vice versa. For this reason (as well as the evident qualitative differences) we conclude that $P_I(k)$ is not a good model for this data.

Q9. Let us call $P_P(k)$ the probability distribution $P(k)$ obtained from the fitted Ising model with pairwise couplings (i.e., with the model $p_{\boldsymbol{g}}(\boldsymbol{s})$ in Equation 9 with the fitted parameters h_i and J_{ij} provided in Canvas). How can you compute $P_P(k)$: can you write the analytical expression of $P_P(k)$ as a function of $p_{\boldsymbol{g}}(\boldsymbol{s})$? Can you plot $P_P(k)$ as a function of k and compare the curve to $P_D(k)$ and $P_I(k)$ previously obtained? (see Figure 16.A of the paper). Which conclusions can you draw?

Figure 14 compares the distributions of observed microstates in the data (orange) with the distributions obtained under the independent-judges assumption (blue) and the fit pairwise Ising model (green). We immediately observe that the pairwise Ising model fits the expected distibution much better than the independent-voting model. Unlike the independent voting model, the Ising model manages to accurately capture the observed peaks at $k = 0$ and $k = 9$, as well as the local (but smaller) maxima at $k = 5$, and the troughs between these regions.

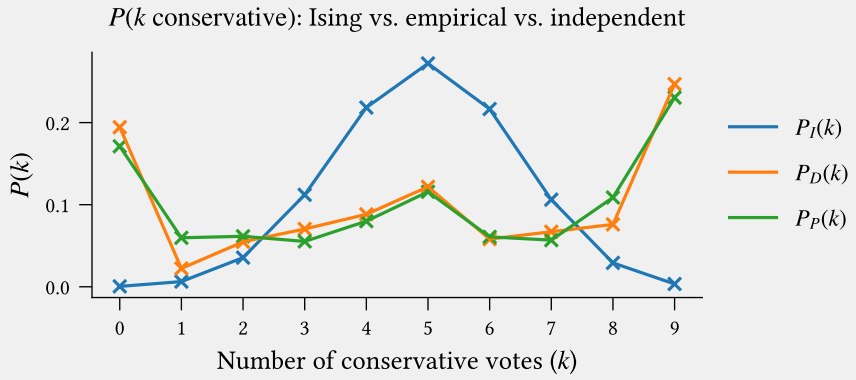


Figure 14: Comparison between the distribution of microstates observed in the data (orange) with that of the fit pairwise Ising model (green) and a model which assumes independent votes between judges (blue).

The three peaks reflect three "stable" states in this system. In a typical case we expect to see either broad agreement (consensus) at one of the two peaks, or a split vote along the political boundary. There is a substantial decrease in

probability mass at $k = 1$ and $k = 8$, indicating that highly-skewed, non-consensus states are uncommon – if almost all judges are voting in a particular direction, the final judge tends to vote this way as well. Of the three stable states, roughly 45% of the probability mass lies at the extremes, indicating that such agreement is quite typical, even if it contradicts individual preferences.

It bears to comment further on the high probability found at the extremes ($k = 0$ and $k = 9$). Under the independent-voting model these states have almost no probability of occurring. To see them attain such high likelihood under the Ising model is evidence of the high impact of pairwise interactions in this system.

Finally, we again find that the pairwise Ising model, while close, does not precisely match the distribution observed in the data, particularly at $k = 1$ and $k = 8$. In both cases the pairwise Ising model overestimates the likelihood of observing such states. While this may simply be a facet of the dataset size, or ambiguity in the labelling, it may also indicate the presence of higher-order interactions which are not being captured by our model.