Inferring Inference

Rajkumar Vasudeva Raju (rv12@rice.edu)
Department of Electrical and Computer Engineering, Rice University

Xaq Pitkow (xaq@rice.edu)
Department of Neuroscience, Baylor College of Medicine
Department of Electrical and Computer Engineering, Rice University

Abstract

We describe a framework to infer canonical computations in distributed neural codes. Our method is based on the theory that the brain performs approximate inference by a message-passing algorithm operating on a probabilistic graphical model. We describe an analysis method that aims to identify this algorithm from neural data elicited during perceptual inference tasks. It simultaneously finds interactions between the decoded variables that define the brain's internal model of the world, along with global parameters that define the message-passing algorithm. The latter parameters are canonical, *i.e.* common to all parts of the graphical model regardless of interaction strength, so they generalize to new graphical models. We apply this analysis method to simulated neural recordings from a simple model brain that performs approximate inference using an advanced mean-field method, and indeed successfully recover the true inference algorithm. We conclude by discussing improvements needed to identify more complex message-passing algorithms.

Message-passing in the brain?

In the brain, information about many variables is distributed across populations of neurons, and this information is recoded by the nonlinear transformations that the neurons implement. We hypothesize that these computations perform probabilistic inference over an internal model the brain has learned. However, exact inference in probabilistic models is intractable except in rare, special cases. Many algorithms for approximate inference in probabilistic graphical models are based on iteratively transmitting information about probability distributions along the graph of interactions. These algorithms go by the name of "message-passing" algorithms because the information they convey between nodes is described as messages. They are dynamical systems whose variables represent properties of a probability distribution.

Examples of message-passing algorithms include belief propagation (Pearl, 2014) and expectation propagation (Minka, 2001). Each specific algorithm is defined by how incoming information is combined and how outgoing information is selected, reflecting different choices of local approximations for intractable global computations. To be a well-defined message-passing algorithm, these core operations must be the same, irrespective of what kinds of latent variables are being inferred, or how strongly they interact in the underlying probabilistic graphical model.

Here we present an analysis method to reverse-engineer such a message-passing algorithm from real or synthetic neural data evoked during perceptual inference tasks. The method simultaneously finds the interactions between the decoded variables that define the brain's internal model of the world, and the global hyperparameters that define the message-passing inference algorithm.

We successfully apply this method to a simple inference algorithm from simulated brain data (see below). This validates the general approach, and encourages us to scale up the method to more complex inference algorithms and real brain data to search for canonical computational structure.

Case study: inferring advanced mean field inference

We analyze data from an inference model that estimates marginal probabilities over an Ising model of N binary variables $s \in \{-1,+1\}^N$ drawn from the joint distribution $p(s) \propto \exp\left(s^\top Js + h^\top s\right)$, where h and J are a bias vector and coupling matrix respectively. The approximate inference algorithm we chose, known as the TAP approximation (Thouless, Anderson, & Palmer, 1977), is an advanced mean field method that estimates the marginal probabilities of each variable s_i, $x_i \approx p(s_i > 0)$ according to the dynamics

$$\dot{x}_{it} = -x_{it} + \sigma\left(\sum_j W\left[J_{ij}, x_{it}, x_{jt}\right] x_{jt} + h_{it}\right) \tag{1}$$

where $\sigma(x) = 1/(1+e^{-2x})$. This can be considered to be a nonlinear neural network with activations x_i and synaptic weights W. Each effective synaptic weight $W\left[J_{ij}, x_i, x_j\right]$ depends on the coupling strength J_{ij} in the underlying graphical model, but is also modulated by the pre- and post-synaptic activity x_j and x_i:

$$W\left[J_{ij}, x_i, x_j\right] = J_{ij} + 2J_{ij}^2(1-2x_i)(1-x_j). \tag{2}$$

Given a time series x_t generated by the algorithm in operation, we use gradient-descent-based optimization to jointly fit the nonlinearity and coupling parameters using the assumed polynomial form

$$\hat{W}\left[J_{ij}, x_{it}, x_{jt}\right] = \sum_{a,b,c} G_{abc} J_{ij}^a x_{it}^b x_{jt}^c \tag{3}$$

over indices a, b, c. This parameterization includes the true model as a special case.

We jointly optimized over both the coupling parameters J and the message-passing parameters G simultaneously by

minimizing the squared error between the true and predicted time-dependent inference dynamics,

$$E[G,J|h] = \sum_{t,i}\left(\sum_j \hat{W}\left[J_{ij},x_{it},x_{jt}\right] - \sum_j W\left[J_{ij},x_{it},x_{jt}\right]x_{jt}\right)^2 \quad (4)$$

where the term $\sum_j W\left[J_{ij},x_{it},x_{jt}\right]x_{jt}$ is constructed from the time series x_t as:

$$\sum_j W\left[J_{ij},x_{it},x_{jt}\right]x_{jt} = \sigma^{-1}\left(\frac{x_{it+1}-(1-\lambda)x_{it}}{\lambda}\right) - h_{it} \quad (5)$$

and λ provides low-pass filtering in the discrete-time version of (1). We also assume that the biases h are fully observed, as they function as sensory evidence both for the algorithm and for our estimation of the algorithm. This is a non-convex optimization, with interesting degeneracies in our parameterization. One such degeneracy is that we can scale all J_{ij} globally by any factor β, and then perfectly compensate by changing G_{abc} by β^{-a}. This is equivalent to increasing the interaction energies and inference 'temperature' at the same time. The parameters also allow a constant offset, $\hat{J} = \beta J + \gamma$. The change in coupling weights is perfectly compensated by $\hat{G}$ that satisfy: $\hat{G}_{0bc} + \gamma\hat{G}_{1bc} + \gamma^2\hat{G}_{2bc} = G_{0bc}$, $\beta\hat{G}_{1bc} + 2\beta\gamma\hat{G}_{2bc} = G_{1bc}$ and $\beta^2\hat{G}_{2bc} = G_{2bc}$, when $a \in \{0,1,2\}$. Figure 1 shows the simultaneous inference of the coupling weights and global parameters for an example graphical model, and illustrates these interesting degeneracies in our parameterization.

We also examine the message-passing algorithm in population codes. Unlike the localist representation discussed above, in such a distributed code the relevant inferred parameters are embedded in the responses of many neurons (Raju & Pitkow, 2016) (Figure 2). Despite the change in representation, we find that the information can still be extracted, and the message-passing identified up to the well-understood degeneracies.

In conclusion, despite the high dimensional parameters and complex cost surface, we can successfully infer a family of algorithms that reproduce the canonical inferential dynamics of model brains. This is a crucial first step toward a new theory of computation in the real brain.

References

Minka, T. P. (2001). Expectation propagation for approximate bayesian inference. In *Proceedings of the seventeenth conference on uncertainty in artificial intelligence* (pp. 362–369).

Pearl, J. (2014). *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann.

Raju, R. V., & Pitkow, X. (2016). Inference by reparameterization in neural population codes. In *Advances in neural information processing systems* (pp. 2029–2037).

Thouless, D. J., Anderson, P. W., & Palmer, R. G. (1977). Solution of 'solvable model of a spin glass'. *Philosophical Magazine*, *35*(3), 593–601.

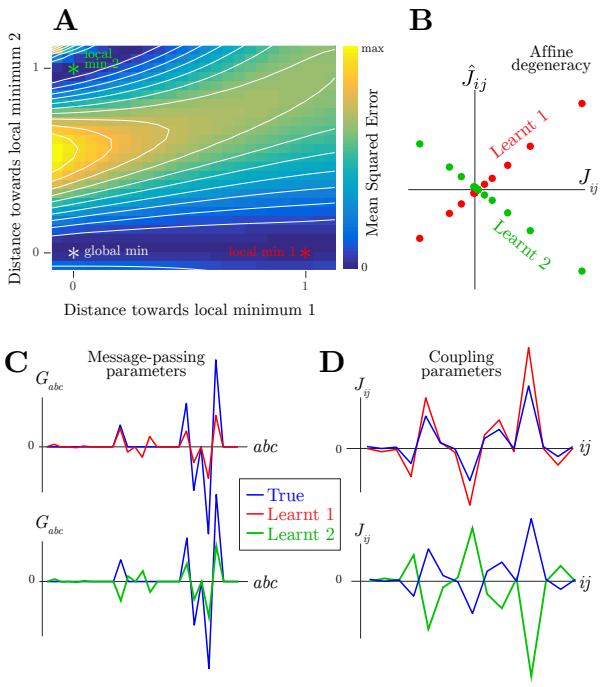

Figure 1: Simultaneous inference of a message-passing algorithm and coupling weights. This example network is a fully-connected 5-node graph with 15 couplings J_{ij}. The true message-passing algorithm can be expressed as a polynomial function with powers up to second order, requiring global coefficients G_{abc} for indices $a,b,c \in \{0,1,2\}$. **A**: Cost function. Shown is a two-dimensional slice through the (3^3+15)-dimensional cost function, with the global minimum defined to be coordinate $(0,0)$ and two local minima defining coordinates $(0,1)$ and $(1,0)$. **B**: The two minima determine two sets of estimated couplings $\hat{J}_{ij}$ (red and green) that are related to the true couplings J_{ij} by affine transformations. **C**,**D**: These two local minima (top and bottom subpanels respectively) provide close matches (red, green) to the message-passing parameters and couplings of the true model (blue). The deviations in the learnt message-passing parameters $\hat{G}$ in panel **C** compensate for these affine transformations in $\hat{J}_{ij}$.

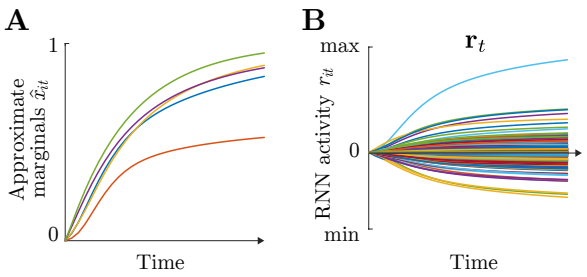

Figure 2: Inferential TAP dynamics (**A**) embedded in simulated population data (**B**). The algorithm's estimated marginals $\hat{x}_t$ can be extracted from the neural data r_t and used to identify the message-passing algorithm (Figure 1).