Supplementary material for

Exploring a Latent Cause Theory of Classical Conditioning

Samuel J. Gershman and Yael Niv
Department of Psychology and Princeton Neuroscience Institute
Princeton University

1 The infinite-capacity mixture model

In this section, we provide more information about the infinite-capacity mixture model described in the main text. Recall our definition of the prior over latent causes:

$$P(c_t = k|\mathbf{c}_{1:t-1}) = \begin{cases} \frac{N_k}{t-1+\alpha} & \text{if } k \text{ is an old cause} \\ \frac{\alpha}{t-1+\alpha} & \text{if } k \text{ is a new cause,} \end{cases} \quad (1)$$

where N_k is the number of observations already generated by cause k (by default it is assumed that $c_1 = 1$).

This distribution over latent causes is known in statistics and machine learning as a *Chinese restaurant process* (Aldous, 1985; Pitman, 2002).[1] Its name comes from the following metaphor: Imagine a Chinese restaurant with an unbounded number of tables (causes). The first customer (trial) enters and sits at the first table. Subsequent customers sit at an occupied table with a probability proportional to how many people are already sitting there, and at a new table with probability proportional to α. Once all the customers are seated, one has a partition of trials into causes. In a Chinese restaurant process mixture model, each cause is linked to a parameterized distribution over features, so that an observation's feature values are determined by its latent cause. Observations generated by the same cause will tend to have similar features by virtue of sharing these parameters.

To gain further intuition for how α governs the number of latent causes, if we were to sample T trials from this distribution, we would obtain on average $\alpha \log T$ unique causes. Note, however, that the posterior over latent causes will not generally obey this law.

[1] The Chinese restaurant process was independently derived by Anderson (1991) in the development of his rational model of categorization.

We assume a Dirichlet distribution as the prior over the multinomial parameters of the observation distribution.[2] This prior expresses the animal's predictions about features in the experiment before any observations have been made. Given that the animal is unlikely to have strong *a priori* predictions about the experiment before it has begun, we parametrized the Dirichlet distribution so that all possible multinomial parameters have equal probability under the prior.

2 Particle filter algorithm

Recall that for trials $1 \dots t$ the vector $\mathbf{c}_{1:t}$ denotes a partition of the trials into clusters and $\mathbf{F}_{1:t}$ denotes the observations for these trials. Our posterior approximation consists of m "particles," each corresponding to a hypothetical partition. In our implementation,[3] the particles are generated by drawing m samples from the following distribution:

$$P(c_t^{(l)} = k) = \frac{1}{m} \sum_{l=1}^{m} P(c_t^{(l)} = k | \mathbf{c}_{1:t-1}^{(l)}, \mathbf{F}_{1:t}), \tag{2}$$

where $c_t^{(l)}$ denotes the latent cause for trial t in particle l, and

$$P(c_t^{(l)} = k | \mathbf{c}_{1:t-1}^{(l)}, \mathbf{F}_{1:t}) = \frac{P(c_t^{(l)} = k | \mathbf{c}_{1:t-1}^{(l)}) \prod_{i=1}^{D} P(f_{t,i} | c_t^{(l)} = k, \mathbf{c}_{1:t-1}^{(l)}, \mathbf{F}_{1:t-1})}{\sum_j P(c_t^{(l)} = j | \mathbf{c}_{1:t-1}^{(l)}) \prod_{i=1}^{D} P(f_{t,i} | c_t^{(l)} = j, \mathbf{c}_{1:t-1}^{(l)}, \mathbf{F}_{1:t-1})}. \tag{3}$$

The first term in Eq. 3 is the latent cause prior (Eq. 1). By default it is assumed that $c_1^{(l)} = 1$. The second term in Eq. 3 is the likelihood of the observed features on trial t given a hypothetical partition and the previous observations. Using a standard calculation for the Dirichlet-Multinomial model (Gelman et al., 2004), we can analytically integrate out the multinomial parameters ϕ associated with each cause to obtain the following expression for the likelihood:

$$\begin{aligned} P(f_{t,i} = j | c_t^{(l)} = k, \mathbf{c}_{1:t-1}^{(l)}, \mathbf{F}_{1:t-1}) &= \int_{\phi} P(f_{t,i} = j | c_t^{(l)} = k, \mathbf{c}_{1:t-1}^{(l)}, \mathbf{F}_{1:t-1}, \phi) P(\phi)\, d\phi \\ &= \frac{N_{i,j,k}^{(l)} + 1}{\sum_j (N_{i,j,k}^{(l)} + 1)}, \end{aligned} \tag{4}$$

where $N_{i,j,k}^{(l)}$ is the number of previous observations with value j on feature i that were generated by cause k in particle l (note that $N_{i,j,k}^{(l)}$ depends on $\mathbf{F}_{1:t-1}$).

[2] The Dirichlet distribution is the conjugate prior for the multinomial distribution, meaning that under this prior the posterior is also a Dirichlet distribution.

[3] This implementation differs slightly from the one described in Gershman et al. (2010). In particular, we use a proposal distribution in this paper that is optimal in the sense that it minimizes the variance of the estimator (Doucet et al., 2001).

The posterior over partitions is then approximated by an average of delta functions placed at the particles:

$$P(\mathbf{c}_{1:t} = \mathbf{c}|\mathbf{F}_{1:t}) \approx \frac{1}{m}\sum_{l=1}^{m} \delta\left[\mathbf{c}_{1:t}^{(l)}, \mathbf{c}\right], \tag{5}$$

where $\delta[\cdot,\cdot]$ is 1 when its arguments are equal and 0 otherwise. As $m \to \infty$ this approximation converges to the true posterior. Although not immediately evident in these equations, learning occurs through maintaining and updating the sufficient statistics of each cluster, namely the cluster-feature co-occurence counts (encoded by $N_{i,j,k}^{(l)}$).

Two things should be noted about this algorithm. First, hypothetical partitions are more likely to the extent that observations assigned to the same cluster are similar; this can be seen in Eq. 4. Second, the features interact multiplicatively in Eq. 3: a partition is more likely to the extent that *all* the observed features are likely under the particle's partition.

The probability of a US for a test observation (i.e., a feature vector in which the US feature is treated as missing data), which we denote by V_t, is calculated according to:

$$\begin{aligned} V_t &= P(f_{t,1} = \texttt{US}|\mathbf{f}_{t,2:D}, \mathbf{F}_{1:t-1}) \\ &= \sum_{\mathbf{c}_{1:t}} P(f_{t,1} = \texttt{US}|c_t, \mathbf{c}_{1:t-1}, \mathbf{f}_{1:t-1,1}) P(c_t|\mathbf{f}_{t,2:D}, \mathbf{F}_{1:t-1,2:D}, \mathbf{c}_{1:t-1}) P(\mathbf{c}_{1:t-1}|\mathbf{F}_{1:t-1}) \\ &\approx \frac{1}{m}\sum_{l=1}^{m}\sum_{k} r_{tk}^{(l)} P(f_{t,1} = \texttt{US}|c_t^{(l)} = k, \mathbf{c}_{1:t-1}^{(l)}, \mathbf{f}_{1:t-1,1}), \end{aligned} \tag{6}$$

where

$$r_{tk}^{(l)} = \frac{P(c_t^{(l)} = k|\mathbf{c}_{1:t-1}) \prod_{i=2}^{D} P(f_{t,i}|\mathbf{F}_{1:t-1}, \mathbf{c}_{1:t-1}^{(l)}, c_t^{(l)} = k)}{\sum_j P(c_t^{(l)} = j|\mathbf{c}_{1:t-1}) \prod_{i=2}^{D} P(f_{t,i}|\mathbf{F}_{1:t-1}, \mathbf{c}_{1:t-1}^{(l)}, c_t^{(l)} = j)}, \tag{7}$$

which is just Eq. 3 excluding the US feature in calculating the cluster assignment probability.

References

Aldous, D. (1985). Exchangeability and related topics. In *École d'Été de Probabilités de Saint-Flour XIII*, pages 1–198. Springer, Berlin.

Anderson, J. (1991). The adaptive nature of human categorization. *Psychological Review*, 98(3):409–429.

Doucet, A., De Freitas, N., and Gordon, N. (2001). *Sequential Monte Carlo Methods in Practice*. Springer Verlag.

Gelman, A., Carlin, J., Stern, H., and Rubin, D. (2004). *Bayesian Data Analysis*. Chapman and Hall/CRC.

Gershman, S., Blei, D., and Niv, Y. (2010). Context, learning, and extinction. *Psychological Review*, 117(1):197–210.

Pitman, J. (2002). *Combinatorial Stochastic Processes*. Notes for Saint Flour Summer School. Techincal Report 621, Dept. Statistics, UC Berkeley.