Supplementary materials to “Anticipating changes: Adaptation and extrapolation in category learning”

Daniel J. Navarro
School of Psychology
University of Adelaide

Amy Perfors
School of Psychology
University of Adelaide

This note contains the supplementary materials to

> Navarro, D. J. & Perfors A. (submitted). Anticipating changes: Adaptation and extrapolation in category learning. Submitted to *Proceedings of the 34th Annual Conference of the Cognitive Science Society.*

It consists of the following sections:

- Proportional shift is equivalent to exponential filtering.
- The six models used in part 1 of the paper
- The five models used in part 2 of the paper

Proportional shift is equivalent to exponential filtering

Footnote 4 in the paper presents an equation that indicates that a simple proportional updating rule corresponds to an exponential filtering scheme. In this first section, we present the derivation of this equation. This result is not new: it is widely known in the statistics literature, but is reproduced here for convenience. The proportional update rule asserts that a prototype previously located at μ_{t-1}, after observing a new item located at x_t is updated as follows:

$$\mu_t = (1-\phi)x_t + \phi\mu_{t-1} \tag{1}$$

where ϕ governs the extent to which the prototype is moved. Since μ_{t-1} itself is the result of an identical updating procredure, we can recursively substitute this updating rule, yielding the following:

$$\begin{aligned}
\mu_t &= (1-\phi)x_t + \phi\mu_{t-1} \\
&= (1-\phi)x_t + \phi((1-\phi)x_{t-1} + \phi\mu_{t-2}) \\
&= (1-\phi)x_t + (1-\phi)\phi x_{t-1} + \phi^2\mu_{t-2} \\
&= (1-\phi)x_t + (1-\phi)\phi x_{t-1} + (1-\phi)\phi^2 x_{t-2} + \phi^3\mu_{t-3} \\
&= (1-\phi)\sum_{i=1}^{t}\phi^{t-i}x_i + \phi^t\mu_0 \qquad (2)
\end{aligned}$$

where μ_0 is initial location of the prototype. If we now let $\tau = -\ln z$, this gives

$$\mu_t = (1 - e^{-\tau}) \sum_{i=1}^{t} e^{-\tau(t-i)} x_i + e^{-\tau t} \mu_0 \tag{3}$$

This is the equation given in Footnote 4 of the paper. In this expression, the $e^{-\tau(t-i)}$ terms assign a weight to each observation, and these decay exponentially with the number of elapsed trials $t - i$. A model that adjusts the category mean by moving it this way will be very order sensitive, since more recent observations are deemed to be more informative about μ_t than old ones.

The six models used in part 1 of the paper

In the first part of the paper, six models are fit to the data set collected by Navarro & Perfors (2009). The description of those models was necessarily abbreviated and imprecise. In this section we present the formal treatment of those models.

Order insensitive models

Formally, all six models can be described as simple Bayesian classifiers. Let x_i denote the stimulus co-ordinate of the ith observed item, and let ℓ_i denote the category label for that item. Similarly, let $\boldsymbol{x}_{1:i} = (x_1, \ldots, x_i)$ denote the collection of stimuli observed so far, and $\boldsymbol{\ell}_{i:i}$ be defined in the same manner for the labels. Then, the classification problem on trial t is to infer ℓ_t the label of the tth item x_t, given all of the previous stimuli $\boldsymbol{x}_{1:t-1}$ and their accompanying labels $\boldsymbol{\ell}_{1:t-1}$. Via Bayes' theorem we can state that the probability that the t-th item belongs to the k-th category is:

$$P(\ell_t = k \mid \boldsymbol{\ell}_{1:t-1}, \boldsymbol{x}_{1:t-1}, x_i) = \frac{P(x_t \mid \boldsymbol{\ell}_{1:t-1}, \boldsymbol{x}_{1:t-1}, \ell_t = k)\ P(\ell_t = k \mid \boldsymbol{\ell}_{1:t-1})}{\sum_k P(x_t \mid \boldsymbol{\ell}_{1:t-1}, \boldsymbol{x}_{1:t-1}, \ell_t = k)\ P(\ell_t = k \mid \boldsymbol{\ell}_{1:t-1})} \tag{4}$$

Since the two categories are equally frequent in this experiment, we may safely assume that $P(\ell_t = k|\boldsymbol{\ell}_{1:t-1}) = 1/2$, and so these terms cancel out. The important quantity that the learner must estimate is $P(x_t \mid \boldsymbol{\ell}_{1:t-1}, \boldsymbol{x}_{1:t-1}, \ell_t = k)$, the probability that the t-th observation would have in fact been x_t if it were generated from category k.

In a standard prototype model, the k-th category is associated with a simple distribution, typically a Gaussian distribution which is characterized by a mean μ_k and a standard deviation σ_k.

$$P(x_t \mid \boldsymbol{\ell}_{1:t-1}, \boldsymbol{x}_{1:t-1}, \ell_t = k) = \frac{1}{\sqrt{2\pi}\sigma_{kt}} \exp\left(-\frac{1}{2\sigma_{kt}^2}(x_t - \mu_{kt})^2\right) \tag{5}$$

Since the mean is a description of the central tendency of the category, it is generally interpreted as corresponding to the category prototype. The estimate of the mean on trial t is denoted μ_{kt} and is constructed by taking the mean of all category members observed to date:

$$\mu_{kt} = \frac{1}{n_{kt}} \sum_{i<t|\ell_i=k} x_i. \tag{6}$$

where $n_{kt} = \sum_{i<t|\ell_i=k} 1$ counts the number of stimuli from category k that have been observed so far. Estimating the standard deviation can be done in one of two ways. If the categories are assumed to be of equal variance, them the same "pooled" estimator σ_t is used for all categories:

$$\sigma_t = \sqrt{\frac{1}{t-1}\sum_{i=1}^{t-1}(x_i - \mu'_i)^2} \tag{7}$$

where μ'_i equals μ_{kt} if $\ell_i = k$. If categories may have unequal variances then each category has its own variance estimate on trial t,

$$\sigma_{kt} = \sqrt{\frac{1}{n_{kt}}\sum_{i<t|\ell_i=k}(x_i - \mu_{kt})^2} \tag{8}$$

Note that since the model estimates all means and variances from the data, there are no free parameters that need to be estimated by fitting to human data.

In a standard exemplar model, the learner does not make any strong assumptions about the form of the category distribution, and instead employs a "kernel density" estimate of the distribution. This gives us the probability distribution:

$$P(x_t \mid \boldsymbol{\ell}_{1:t-1}, \boldsymbol{x}_{1:t-1}, \ell_t = k) = \frac{1}{\lambda n_{kt}}\sum_{i<t|\ell_i=k}\exp(-\lambda|x_t - x_i|) \tag{9}$$

Since the kernel width parameter λ is not estimated by the model itself, there is one free parameter in this model.

Introducing order sensitivity

None of these models are order-sensitive, since the underlying statistical model in all cases assumes that data are generated independently from a single fixed category distribution. A simple way to address this is to suppose that the category mean moves in an unsystematic way from trial to trial. If so, older observations are less informative as to the category distribution as newer ones. For instance, if the true category mean changes according to a standard autoregressive process (specifically AR(1)), then this decay will be exponential in character. This suggests a simple exponentially-weighted estimate:

$$\mu_{kt} = \frac{\sum_{i<t|\ell_i=k} w_i x_i}{\sum_{i<t|\ell_i=k} w_i} \tag{10}$$

where $w_i = \exp(-\tau(t-i))$ is an exponentially decaying function of recency, and the decay rate τ is a parameter that must be estimated, and can be mapped onto the ϕ parameter in a proportional updating rule using the result discussed at the start of the note, where $\phi = \exp(-\tau)$. Since the ϕ parameterization is simpler, we report all results in terms of ϕ

No new parameters are required to construct the variance estimates, since we assume that the variance does not change. Instead, the variance estimators differ only insofar as they must now accommodate the variation in the mean,

$$\sigma_{kt} = \sqrt{\frac{1}{n_{kt}}\sum_{i=1}^{t-1}(x_i - \mu_{ki})^2} \tag{11}$$

with a similar alteration applying for the equal variance case. Strictly speaking, the estimators described above are slightly suboptimal, in the sense that the variance estimates could be improved by "looking backwards" and using the benefit of hindsight to construct a better estimate of an old mean estimate μ_{ki} in light of new data. However, the simpler version is more than sufficient for current purposes, so we avoid introducing any fancier estimates.

In any case, by analogy we can construct a time-sensitive version of the exemplar model, simply by weighting each exemplar by recency. The category distribution now becomes:

$$P(x_t \mid \boldsymbol{\ell}_{1:t-1}, \boldsymbol{x}_{1:t-1}, \ell_t = k) = \frac{\sum_{i<t|\ell_i=k} w_i \exp(-\lambda |x_t - x_i|)}{\lambda n_{kt} \sum_{i<t|\ell_i=k} w_i} \tag{12}$$

where again $w_i = \exp(-\tau(t-i))$. Note that there are two separate "generalization gradient" parameters now: λ governs the generalization across psychological space, and τ governs the generalization across different time points.

The five models used in part 2 of the paper

Part 2 of the paper applies five models to a new experiment, and again all models can be folded into the Bayesian classification framework discussed in the previous section. Because the categories in the new experiment are unimodal and have equal variance, all five models are based on the equal variance prototype model. Two of the five models have already been discussed: they are the equal variance prototype models with and without the exponential weighting scheme (referred to as "standard" and "recency" in the paper). The third model, referred to as "recency + bias", introduces a constant correction term β for the location of the prototype. This is

$$\mu_{kt} = \beta + \frac{\sum_{i<t|\ell_i=k} w_i x_i}{\sum_{i<t|\ell_i=k} w_i} \tag{13}$$

The fourth and fifth models both rely on a linear regression approach to adapt the category mean. After having observed the data from the first $t-1$ trials, the learner has seen some number of stimuli that belong to the category. Each such observation can be expressed as the pair (x_i, i), where x_i is the stimulus and i is the trial number on which the i-th stimulus was observed. Note that for the k-th category, we use only the stimuli and trial numbers for items that belong to category k. Using these data, the learner estimates a regression model of the form:

$$x_i = b_1^{(kt)} i + b_0^{(kt)} + \epsilon_i \tag{14}$$

where ϵ_i is the residual associated with the i-th trial. The superscripts in $b_1^{(kt)}$ and $b_0^{(kt)}$ indicate that these are the regression coefficients the learner has estimated for the k-th category after having observed the first $t-1$ trials. The corresponding estimate for the prototype μ_{kt} is

$$\mu_{kt} = b_1^{(kt)} t + b_0^{(kt)} \tag{15}$$

The difference between models four and five lies in how the regression coefficients are calculated. In model four, the learner estimates the $b_1^{(kt)}$ and $b_0^{(kt)}$ using a standard, unweighted linear regression, treating each observations as equally important. Model five weights more recent observations more heavily, using the exponential weighting scheme discussed previously. (Formally, the latter model requires us to make use of the weighted covariance between the stimulus representation and the trial number: this is not difficult, but we omit the details for the sake of brevity).

For all models used in this experiment the variance estimates for the category are calculated using the equal variance version (i.e. pooled estimate) of Equation 11. The only differences between the variance estimates lies in the fact that each model produces different estimates for the locations of the prototype.