Are Mixtures-of-Experts Psychologically Plausible?

Sébastien Hélie[1], Gyslain Giguère[1], Denis Cousineau[2], and Robert Proulx[3]

[1] Université du Québec À Montréal, Computer Science, C.P. 8888 Succ. Centre-Ville,
Montréal, H3C 3P8, Canada
{Helie.Sebastien, Giguere.Gyslain}@courrier.uqam.ca
[2] Université de Montréal, Psychology, C.P. 6128 Succ. Centre-Ville, Montréal, H3C 3J7,
Canada
Denis.Cousineau@umontreal.ca
[3] Université du Québec À Montréal, Psychology, C.P. 8888 Succ. Centre-Ville, Montréal,
H3C 3P8, Canada
Proulx.Robert@uqam.ca

In this paper, we explore the psychological plausibility of mixture-of-experts models. This type of models is referred to in psychology as the knowledge partitioning theory (KP). Over the years, it was argued that: 1) KP is a necessary part of function learning, 2) the usefulness of KP is proportional to task difficulty, and 3) the experts used by humans to perform function learning tasks are always linear. In the present study, these statements were tested by modifying the test display. The results show that increasing the difficulty of stimulus estimation unexpectedly resulted in non-linear KP. Also, adding less useful information to the display resulted in a smaller proportion of partitioning participants. We conclude that mixture-of-experts are adequate psychological models for KP, but that the linearity and ubiquity claims need to be weakened.

1 Introduction

The main goal of any intelligent agent is to adapt to its environment. This is often accomplished by finding contextual cues which are informative about the action to be performed next. One way of achieving such adaptation is to use several processes (or *experts*), each associated to a particular context. If enough experts are available to adequately cover the entire space of action, no search is necessary: a network can be trained to gate each situation to the correct expert, which computes the best action according to context. This type of architecture is called a mixture-of-experts [1, 2]. This class of models is particularly effective in situations where different (even contradictory) responses are appropriate according to situations [2] (such as multispeaker vowel recognition [1]).

One task accomplished by humans in order to adapt to their ever changing environment is categorization. In the particular case where exemplars (inputs) and categories (outputs) are continuous, one usually extracts the function relating the input to the output (function learning), instead of performing standard associative learning. Simple examples of function learning includes estimating the distance of a moving object according to its size on the retina, or how long you can stay in the sun before you burn.

1.2 Function learning and its application to forest fires

Another situation where learning a function is necessary is when one must estimate the speed of spread of forest fires [3, 4]. When the slope of the terrain and the wind direction are in opposition, a forest fire spreads uphill at a speed negatively related to wind speed, unless the wind becomes strong enough to overcome the fire's natural propensity to spread uphill. From this moment on, the fire will spread downhill at a speed positively related to the increasing wind speed. Overall, the function relating speed of spread to wind speed is a concave quadratic function where the vertex indicates the point at which the force applied by the wind overpowers the tendency of the fire to spread uphill.

Another important aspect of firefighting is the use of back-burning fires to control the reach of the to-be-controlled fire. Back-burning fires are lit and managed by firefighters to starve the to-be-controlled fire of fuel. Usually, a back-burner is lit when the wind speed is low; otherwise the firefighters might lose control of this second fire.

The type of fire (back-burning, to-be-controlled) is an important cue which facilitates the estimation of a fire's spreading speed: instead of considering a quadratic function to determine the propagation of the fire, firefighters can use two linear functions: a decreasing function associated to the back-burning context and an increasing one associated to the to-be-controlled context. This two-stage decision process (cue identification and response selection) is equivalent to a mixture-of-experts architecture [1, 2] which identifies the context first and uses a different, linear, expert accordingly. In [3], experienced firefighters were shown to use this two-stage strategy. Lewandowsky and Kirsner argued that the association between the context (type of fire) and the linear functions had been learnt through their many years of experience. Accordingly, it was argued that this two-stage process was ubiquitous in expertise: this theory was called knowledge partitioning (KP).

To test the KP theory, another experiment was designed to assess whether novices would also use this strategy in a function learning task [4]. In this second experiment, the participants were taught basic firefighting background knowledge and trained in a standard function learning experiment using a concave quadratic function. Every stimulus (wind speed) was also accompanied by a context label, which was systematically associated to a different half of the function during training. This manipulation aimed at recreating the bias present in experienced firefighters' knowledge, for which back-burners are usually encountered in low wind speed situations and to-be-controlled fires in high wind speed conditions. At test, every stimulus was presented twice: once as a back-burner and once as a to-be-controlled fire. Results showed that participants easily achieved the task but, more importantly, that while spreading speeds were almost perfectly estimated when wind speeds appeared in their usual context, they were systematically underestimated when shown in the unusual context. Hence, participants' gave dramatically different responses to identical stimuli presented in different contexts, which supports the KP theory.

This support for the KP theory constitutes empirical evidence in favor of the psychological plausibility of mixture-of-experts models [1, 2]. In particular, one such model was proposed to explain human performance: POpulation of Linear Experts (POLE) [5]. In POLE, when a stimulus is encountered, a gating mechanism directs it to the correct expert, which represents one of many linear functions with different slopes and intercepts. There are enough experts to cover the entire stimulus space and only the gating system has adjustable weights. Once an expert is chosen, it computes the answer accordingly.

This model [5] possesses three important properties. First, POLE accounts for all past results in the function learning literature by using KP. Second, experts do not blend together. Therefore, the system always commits to a cue-value and chooses an expert accordingly (KP). Third, each expert represents a linear relationship between the stimulus and the response. These properties and the preceding empirical results were used to make certain claims concerning the generality and properties of KP [4, 5]: 1) KP is always used when the association between the context and a part of the function is systematic, 2) the usefulness of KP is proportional to task difficulty, and 3) humans always partition into *linear* subfunctions to achieve complex function learning tasks.

In the present study, we empirically tested the preceding claims related to KP in general [3, 4] and to POLE in particular [5]. Human data were collected by altering Lewandowsky et al.'s [4] experimental settings. A first group performed the same task as Lewandowsky et al. [4]. A second group was trained in the same task with smaller stimuli: this task was hypothesized to be more difficult and should lead to an equal or higher proportion of partitioners. The third group was trained with settings identical to the second, except that information was added to the display (constant visual markers). Because the necessary conditions for finding KP were present in this condition (e.g. a systematic association between particular values and a context), we should find as many partitioners as in the small stimuli condition.

2 Experiment

2.1 Method[1]

2.1.1 Participants

Fifty-four undergraduate students from the Université de Montréal participated in this experiment. Eighteen participants were trained in a reproduction of Lewandowsky et al. [4] (control group), eighteen were trained with small stimuli (small stimuli group), and the remaining participants were trained with small stimuli and supplemental

[1] This experiment is an extension of Lewandowsky et al.'s Experiment 1, systematic condition [4]. Therefore, this section bears on their original methodology. However, it was brought to our attention that our participants were given more extensive background knowledge. Nevertheless, performance was not qualitatively different, as shown by the results from the control group.

information (extra information group). In each group, six participants were assigned to the complete condition, six to the left-only condition, and the remaining six to the right-only condition. Participants in the complete conditions received 7$ as compensation for their time, and those in the left-only or right-only conditions received 5$. The experiment was conducted in French.

2.1.2 Material

Participants were tested individually. All instructions and stimuli were presented on 43 cm (17 inch) monitors connected to PCs. Participants were positioned approximately 60 cm away from the monitor. The experimental task was programmed using Sun Microsystems' Java J2SDK1.4.1. The program was used to present the material and record the participants' answers.

2.1.3 Stimuli

Participants were expected to learn a concave quadratic function in which the fire's spreading speed (F) was related to wind speed (W) in the subsequent manner: $F(W) = 24.2 – 1.8W + 0.05W^2$. Wind direction always opposed slope, and the vertex of the function (W = 18) represented the point at which the force of the wind balanced the effect of the slope. To the left of that point, fire speed decreased with increasing wind speed, without changing the direction of the fire spread. Lewandowsky et al. [4] referred to these fires as "slope-driven". To the right of the vertex, fires were "wind-driven" and their speed increased as a function of wind speed. During training, 36 stimuli were used, ranging from wind speeds of 0 to 36, omitting the vertex of the function. At test, the omitted wind speed of 18 was included, resulting in a total of 37 transfer stimuli.

On each trial, a horizontal arrow, whose length was proportional to a particular wind speed (henceforth referred to as the stimulus), was shown at the top of the display. The minimal arrow length, associated with the value 0, was approximately 5.8 cm for the small stimuli and extra information groups and 0.7 cm for the control group. The maximal length, associated with the value 36, was approximately 26 cm for the small stimuli and extra information groups and 31 cm for the control group. Thus, in the small stimuli and the extra information groups, the shortest arrow occupied 1/6 of the display and the longest 5/6. In the control group, the arrows spanned the entire monitor. No numerical values for wind or fire speed were shown. Participants were to consider each fire in a context represented both by a color-coded verbal label and arrow (blue for *Back-burning* and red for *Firefighting*). In the extra information group, visual markers were added to the display to indicate the minimal and maximal possible stimulus lengths. The markers were the only difference between the small stimuli group and the extra information group.

Participants were asked to predict the speed of the fire (notwithstanding its direction of spread) by moving a sliding pointer along a 23.3 cm-scale positioned in the left part of the display. The scale was labeled *slow* at the bottom and *fast* at the top, without any incremental values or tick marks.

After each training trial, the participant's response was followed by a feedback arrow. The arrow was located next to the response scale to indicate the correct speed of spread. Also, a message appeared in a rectangle at the bottom center of the screen to encourage the participant to perform better (yellow rectangle) or to indicate that the response was satisfying (green rectangle). Predictions deviating by 5 or more units (approximately 7.2 cm) from the correct answer were accompanied by the former (yellow message) while acceptable performances were accompanied by the latter (green message). Participants were required to acknowledge feedback by a mouse click. The inter-stimulus interval (ISI) was 2 seconds, and the textual context-label always preceded the stimulus by 1 second. At test, feedback was absent.

2.1.4 Procedure

The procedure was identical for all groups and varied according to conditions. In all conditions, each stimulus was presented five times during training. Hence, there was a total of 180 trials for the complete conditions, but only 90 trials for the left-only and right-only conditions (because training was restrained to one half of the function). In all conditions, 90% of fire speeds occurred in their respective contexts, and the remaining 10% were presented in the opposite context. However, in the left-only and the right-only conditions, all stimuli were presented in the same context (back-burning for left-only and firefighting for right-only). All magnitudes were presented once within each block of 36 trials (18 for the left-only and the right-only conditions), except during the first block, where magnitudes were presented in a blocked manner.

After completion of the training trials, participants in all conditions completed the same transfer test. The transfer test involved predicting the fire speed of all stimuli in both contexts.

2.2 Results

2.2.1 Training

The participants' *Absolute Deviation from Function* (ADF) was used to evaluate learning[2]. Learning curves are shown in Fig. 1. As seen, participants in all conditions from all groups improved their ADF and were thus able to learn the function. Also, Fig. 1 suggests no effects of groups or conditions.

A Group (small stimuli vs. extra information vs. control) × Condition (complete, left-only, right-only) × Block (5, repeated measures) ANOVA was performed on the participants' ADF to corroborate what Fig. 1 hinted. First, the participants were able to diminish their ADF with practice: The mean ADF was 2.33 in the first block and diminished to 1.74 in the fifth block ($F(4, 176) = 14.32$, $p < .01$). However, this effect must be interpreted with care, because the Block × Group interaction was significant ($F(8, 176) = 10.24$, $p < .01$). Thus, the group effect was further decomposed within each block. The groups significantly differed in the first block of training ($F(2, 44) = 28.42$, $p < .01$) but were similar in all other blocks (all $F(2, 44) < 1.63$, $p > .05$). *Tukey A post hoc comparisons* showed that the control group was significantly better than the other two at the beginning of the

[2] The performance of one participant from the small stimuli group, complete condition, deteriorated with practice ($F(4, 175) = 2.54$, $p < .05$). Therefore, this participant was excluded from the following analyses.

task (both differences > 0.96, *p* < .01). However, as suggested by the absence of group effect in the remaining blocks, this difference disappeared with training.

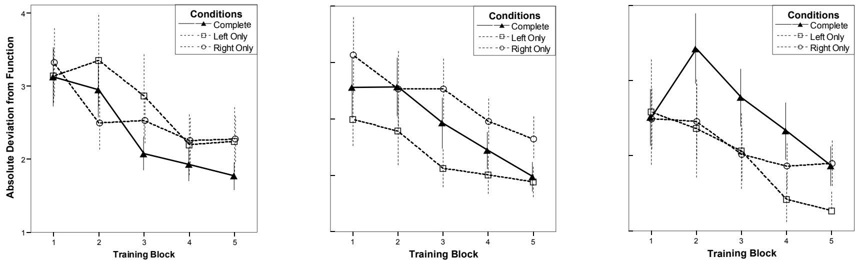

Fig. 1. Participants' ADF during the training phase. The left panel shows performance of participants in the extra information group, the middle panel participants in the small stimuli group and the right panel shows the control group

2.2.2 Group performance at test

KP can be detected experimentally by a difference in responses to a given stimulus in different contexts [4]. Fig. 2 shows transfer performances for participants trained in the complete conditions. As seen in the left panel, participants trained with extra information learned the function quite well. Surprisingly, answers in both contexts matched the quadratic function and were not affected by context.

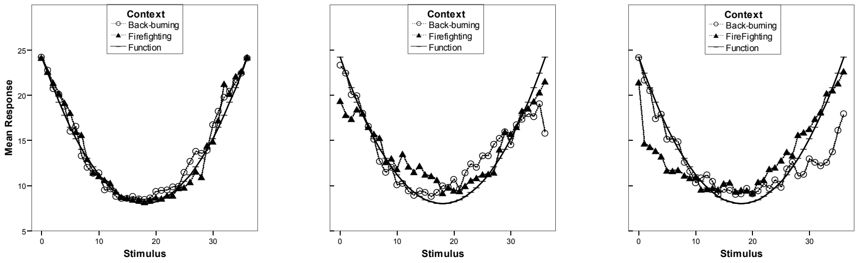

Fig. 2. Mean responses at test in each context. Panels represent the same groups as in Fig. 1

Responses of participants trained with small stimuli are shown in the middle panel. As expected, their responses at test were affected by the context (compare with the left panel), but in an unexpected way. In comparison, the deviations found by Lewandowsky et al. [4] were systematic: low wind speeds resulted in an underestimation of the speed of fire spread in the firefighting context and high wind speeds accordingly resulted in underestimations in the back-burning context. This is exactly the pattern of results found in the control group (see the rightmost panel). In the middle panel, the underestimations are present (to a lesser extent), but mid-range wind speeds were overestimated.

A better way to highlight the difference in responses to a given stimulus is to compute the signed differences [4]. A signed difference is computed by subtracting the answer given at test to each stimulus in the back-burning context from the answer given to the same stimulus in the firefighting context. Signed differences randomly aggregated around zero would suggest the absence of partitioning, while signed differences systematically deviating in one direction would indicate the presence of partitioning.

The left panel of Fig. 3 confirms that participants trained with extra information are not partitioning their knowledge: their signed differences are randomly aggregated around the abscissa. Participants in the control group did show the expected pattern of results: signed differences are negative to the left of the vertex and positive to the right. The signed differences of participants trained with small stimuli are more intriguing (middle panel): they are substantially deviating from the abscissa in a sine-like way.

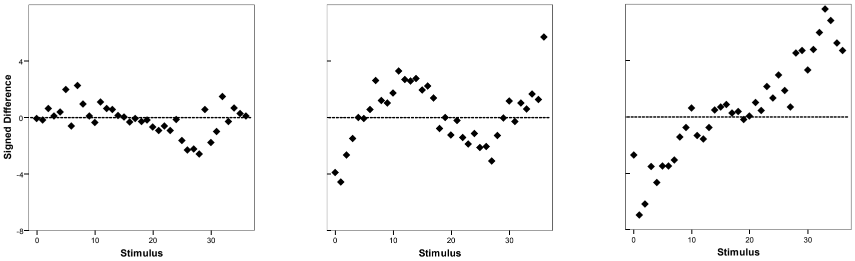

Fig. 3. Signed differences of the participants at test. Panels represent the same groups as in Fig. 1

2.2.3 Individual results at test

Considering that Lewandowsky et al. [4] found important individual differences relating to KP, it is relevant to verify if the effects found in section 2.2.2 were representative of the entire groups of participants. A novel, statistical way of classifying the participants as partitioners (P) or non-partitioners (NP) is to individually plot their signed differences and estimate the best-fitting linear model using a linear regression. A slope which is significantly different from zero suggests a systematic effect of context, namely KP. On the other hand, a slope of zero suggests no clear effect of context. Table 1 shows the slope and intercept individually estimated for each participant.

Table 1 shows that all but one participant fit a model with an absolute slope of 0.05 or less in the group trained with additional information. Because these slopes did not differ significantly from zero (*p* = .05), these participants were classified as NPs. The exact opposite was true of participants in the small stimuli group: All but one participant had an absolute slope greater than 0.20. Hence, these participants were classified as Ps. In the control group, the best-fitting slope of two of the six participants was smaller than 0.10: these slopes did not significantly differ from zero (p = .05) and these participants were classified as NPs. The four remaining participants were classified as Ps.

The proportion of Ps in the small stimuli group significantly differed from the proportion of Ps in the extra information group according to a binomial test ($B(5, 1/6) = 4$, $p < .01$). The proportion of Ps in the small stimuli group (80%) is well in range with past literature[3] while the proportion of Ps in the group trained with extra information (16.7%) is below past results. The proportion of Ps in the control group (67%) is similar to Lewandowsky et al.'s results [4] and does not significantly differ from the small stimuli group ($B(6, 4/5) = 4$, $p > .05$). However, this proportion of Ps differs from the proportion found in the extra information group ($B(6, 1/6) = 4$, $p < .01$).

Table 1. Estimated Parameters for the Best-Fitting Linear Models

		Estimated			
	Participant	Slope	Intercept	r^2	Classification
Extra Information					
	110	-0.03	0.63	0.02	NP
	111	0.00	-0.13	0.00	NP
	112	-0.25	5.48	0.34	P
	120	-0.02	0.37	0.01	NP
	121	0.05	-1.45	0.07	NP
	122	0.01	-1.51	0.00	NP
Small Stimuli					
	210	-0.41	8.03	0.58	P
	211	-0.02	0.19	0.00	NP
	212	0.22	-4.19	0.29	P
	220	0.63	-9.04	0.93	P
	221	-0.23	2.68	0.32	P
Control					
	310	0.60	-10.7	0.87	P
	311	-0.02	0.17	0.01	NP
	312	-0.07	2.16	0.09	NP
	320	0.54	-9.45	0.69	P
	321	0.13	-2.61	0.14	P
	322	0.65	-8.88	0.89	P

Note. P = Partitioners; NP = Non-Partitioners

Together, these results suggest that when extra information is present in the display, fewer participants use the KP heuristic, even if the added information is far less useful than the context. Also, it is noteworthy that all the Ps in the control group showed positive slopes, which is consistent with the linear experts hypothesis [4, 5]. However, half of the Ps in the small stimuli group and the only P in the extra information group had negative slopes, which is consistent with the sine-like pattern of Fig. 3. The overestimation of moderate wind speeds is also present in the middle

[3] Precisely, previous research found between 13% [6] and 50% [5] of participants who were not partitioning their knowledge.

panel of Fig. 2 and further inspection of the middle panel suggests a partitioning of the stimuli in two quadratic functions with skewed vertices. Therefore, diminishing the stimulus' length does not prevent participants from using KP but entails a different, non-linear, type of partitioning, which is not consistent with POLE's predictions [5].

2.2.4 Independence of knowledge parcels

As Lewandowsky et al. [4] pointed out, participants who were uniquely trained on the left or right part of the function represent extreme cases of KP: they possess a single expert, associated with a single context. Therefore, if the knowledge of Ps in each context is truly independent, their responses should be similar to the left-only condition in the back-burning context and the right-only condition in the firefighting context. In the case of non-linear Ps, responses in the back-burning context were similar to responses from participants uniquely trained in this particular context (left-only condition: $r = 0.87$). However, the correlation between partitioners' responses in the firefighting context and those from the right-only condition was smaller ($r = 0.69$). This difference is significant according to Fisher's Z transform test ($Z = 2$, $p < .05$). Therefore, the back-burning parcel of knowledge seems more hermetic than the firefighting parcel. Also, results from Lewandowsky et al. suggested higher correlation coefficients [4].

In the case of linear Ps, responses from knowledge partitioners were similar to responses from participants trained in the left-only ($r = 0.81$) and right-only ($r = 0.83$) conditions (in the back-burning and firefighting contexts respectively). Also, the difference between correlation coefficients is not statistically significant ($Z = 0.25$, $p > .05$). Knowledge about the other half of the function acquired in another context did not affect the participants' responses, suggesting that knowledge was completely partitioned.

3 General discussion

In the Experiment, the usual settings used to assess the presence of KP [4, 5] were varied to check the robustness of this phenomenon. Precisely, two modifications were made: reducing the stimulus range, and adding potentially distracting information. First, results from the control group confirmed the adequacy of our reproduction of [4]. Second, it is well established that diminishing the span of the stimuli increases discrimination difficulty [7], hence making stimulus estimation more difficult. In the small stimuli group, participants, classified using our novel method, used KP in an expected proportion, but showed negatively-sloped best-fitting linear models (Table 1). This counter-intuitive result was first hinted by sine-like signed differences (Fig. 3) and the use of non-linear expert functions with skewed vertices (Fig. 2). Finally, adding less useful information to the display was sufficient to prevent most participants from using KP to achieve the task, even if the sufficient conditions for the use of this strategy were present. However, these participants, who did not use KP to simplify the function, were still able to learn it (as shown by an absence of group effect in the ANOVA).

3.1 Implications for current cognitive modeling

Our findings have numerous implications for cognitive modeling. In particular, results from the small stimuli group are challenging the POLE model [5]: when stimuli are more difficult to estimate, participants still partition their knowledge but non-linear experts are used. This can be explained by the added difficulty in the estimation of the function's vertex: if the range of applicability of an expert is fuzzy, the other experts must try to compensate. This strategy is adaptive because, by using more complex functions, the error resulting from an erroneous choice of expert is minimized. Hence, the results from the small stimuli group, while challenging to POLE's predictions, do not invalidate mixture-of-experts models in general [1, 2].

The results from the extra information group are more problematic to both POLE [5] and general mixture-of-experts models [1, 2], because they show that when potentially distracting information is present in the display, participants do not seem to be using the KP heuristic. Instead, participants are learning the quadratic function by simple associative learning. These findings might still be explained by the degenerate case of the mixture-of-experts, in which a single quadratic expert is used.

Together, these results confirms that KP [3-6], which is the empirical counterpart to mixture-of-experts models [1, 2], is a strategy used to achieve psychological tasks. However, this heuristic is less ubiquitous than Lewandowsky and his colleagues previously thought [5], and the constraint of using linear experts is too restrictive. Therefore, mixture-of-experts are adequate models of human cognition but further research is needed to detect the presence of experts (to distinguish simple associative learning from the degenerate case of using a single expert) as well as to determine the nature of the experts used to achieve particular tasks.

References

1. Jacobs, R.A., Jordan, M.I., Nowlan, S.J., Hinton, G.E.: Adaptive Mixtures of Local Experts. Neural Computation **3** (1991) 79-87
2. Bishop, C.M.: Neural Networks for Pattern Recognition. Oxford University Press, New York (1995)
3. Lewandowsky, S., Kirsner, K.: Knowledge Partitioning: Context-Dependent Use of Expertise. Memory & Cognition **28** (2000) 295-305
4. Lewandowsky, S., Kalish, M., Ngang, S. K.: Simplified Learning in Complex Situations: Knowledge Partitionning in Function Learning. Journal of Experimental Psychology: General **131** (2002) 163-193
5. Kalish, M.L., Lewandowsky, S., Kruschke, J.K.: Population of Linear Experts: Knowledge Partitioning and Function Learning. Psychological Review **111** (2004) 1072-1099
6. Yang, L.-X., Lewandowsky, S.: Context-Gated Knowledge Partitioning in Categorization. Journal of Experimental Psychology: Learning, Memory, and Cognition **29** (2003) 663-679
7. Goldstein, E.B.: Sensation & Perception. 5th edn. Brooks/Cole Publishing Company, Pacific Grove (1999)