

Bayesian Approaches to Color Category Learning

Thomas L. Griffiths*

Department of Psychology, University of California at Berkeley, Berkeley, CA, USA

Synonyms

[Ideal observer models of color category learning](#); [Rational models of color category learning](#)

Definition

Bayesian approaches to color category learning formalize learning as a problem of Bayesian inference, requiring the learner to form generalizations that go beyond observed examples of members of a category. This formal framework can be used to make predictions about both individual judgments and how populations form color categories.

Color Category Learning

One of the challenges that children face as they acquire a language is discovering how words are used to refer to different colors. While human languages demonstrate variation in how they partition the space of colors, there are also clear regularities in the kinds of systems of color categories that are used [1]. This raises two important questions: How might color categories be learned? And how might regularities in systems of color categories across languages be explained?

Learning color categories is an inductive problem, requiring learners to make an inference from labeled examples of colors to a full system of color categories. As in other domains of perception [2], an “ideal observer” model can be used to explore the optimal solution to this problem. Let h denote a hypothesis about a possible system of color categories and d the observed data – a set of labeled examples (such as “This color is *blue*, and this color is *yellow*”). If learners represent the degree of belief in the truth of each hypothesis with a probability, $P(h)$, then the ideal solution to the problem of updating these beliefs in light of the data d is provided by Bayes’ rule:

$$P(h|d) = \frac{P(d|h)P(h)}{\sum_{h'} P(d|h')P(h')}$$

where $P(h|d)$ (known as the *posterior* probability, in contrast to the *prior* probability $P(h)$) indicates the degree of belief assigned to h after observing d and $P(d|h)$ (known as the *likelihood*) indicates the probability of seeing d if h were true.

The sum in the denominator of Bayes’ rule ranges over all possible hypotheses and ensures that $P(h|d)$ is a valid probability distribution, summing to 1. The key idea behind Bayes’ rule can be obtained by ignoring this constant and simply inspecting the numerator: The new beliefs of the learner result from combining the previous beliefs, captured in the prior distribution $P(h)$, with the probability of the

*Email: tom_griffiths@berkeley.edu

observed data under each hypothesis, expressed by the likelihood $P(d|h)$. The prior distribution captures the expectations of the learner, but also indicates which hypotheses are easy or hard to learn. A hypothesis that has low prior probability requires stronger evidence (in the form of a higher likelihood) to end up with a high posterior probability and so will be harder to learn. The prior distribution thus provides a way of encoding the perceptual or learning biases of the learner, favoring some hypotheses over others.

The Bayesian approach to modeling learning has proven successful in accounting for human behavior in a wide range of tasks [3]. In particular, Bayesian models have been used to account for how people learn new concepts and new words. Tenenbaum and Griffiths [4] presented an account of how people form generalizations from examples, such as inferring what other numbers might belong to a set when told that the set contains 2, 8, and 64. Under this account, hypotheses correspond to possible sets of numbers, and the likelihood is obtained by calculating how likely it is that the examples would be observed if they were sampled at random from this set. Xu and Tenenbaum [5] showed that a closely related model captured how children learned nouns corresponding to sets of objects, such as determining the appropriate referent of words corresponding to “Dalmatian” or “dog.” These results suggest that a Bayesian approach might also be fruitful for explaining the acquisition of terms for color categories.

Dowman [6] presented a model that took exactly this approach, providing a Bayesian account of color category learning. In this model, the space of colors is reduced to a one-dimensional ring of hues. Each color category then corresponds to an interval on this ring, picking out a set of adjacent colors. Labeled examples of color categories are assumed to be sampled from the categories at random, with a small probability of an error taking place. This makes it possible to calculate the probability of any observed set of labeled examples for each candidate interval from which they might be drawn, providing the likelihood $P(d|h)$. Bayes’ rule can then be used to compute a posterior distribution over possible intervals for each color category. The probability that a color that has not previously been labeled belongs to that category is then obtained by summing the probability of all intervals that contain that color under the posterior distribution. Dowman demonstrated that this model made reasonable inferences for simplified versions of the systems of color categories from real languages, such as Urdu.

The predictions that Dowman’s model makes about learning of color categories have not been directly tested with human learners, but results in other domains and with other species provide support for this approach. As noted above, Xu and Tenenbaum [5] found that a very similar model accounted well for the generalizations that children made in learning novel words describing sets of objects. In addition, Jones, Osorio, and Baddeley [7] found that poultry chicks form generalizations about colors in a conditioning task that are consistent with a model based on that of Tenenbaum and Griffiths [4].

The results summarized so far indicate how Bayesian models might be used to explain learning of color categories. The same models also have the potential to provide insight into why regularities exist in the systems of color categories that appear across human languages. Dowman [6] explicitly had this goal in mind in defining his Bayesian learning model, which was used as a component in a simulation of the cultural transmission of systems of color categories. Dowman’s aim was to investigate the consequences of cultural transmission of systems of color categories among a set of agents that used a realistic approximation to human learning. He found that cultural transmission by Bayesian agents produced systems of color categories with properties similar to those seen across human languages, providing a potential explanation for the source of those regularities.

Dowman’s simulation of cultural transmission was an instance of a more general approach to exploring the origins of different kinds of structure in human languages, known as iterated learning [8]. In an iterated learning model, a set of agents each learn from some observed data and generate data that is observed by other agents. The simplest case is where the agents form a chain, with each agent learning from data generated by the previous agent and generating data that is provided to the next agent. Griffiths and Kalish [9] showed that when this form of iterated learning is carried out by Bayesian agents who all have the same

prior, the hypotheses considered by those agents eventually converge to a distribution that matches the prior distribution. More precisely, the probability that an agent selects a hypothesis h converges to the prior probability of that hypothesis $P(h)$ as the chain gets longer.

The results of Dowman [6] and Griffiths and Kalish [9] raise an interesting question: Can the regularities seen in systems of color categories across human languages be accounted for by cultural transmission producing convergence on a shared prior distribution? To explore this question, Xu, Dowman, and Griffiths [10] conducted an experiment in which human participants simulated cultural transmission by iterated learning. Each participant was given some examples of colors from novel categories and then asked to generalize to a larger set of colors. The generalization responses were then used to generate the examples that were seen by the next participant. Over time, the systems of color categories produced by the participants converged to forms that were consistent with the regularities seen across human languages. These results support the idea that cultural transmission and perceptual or learning biases of the kind that might be captured by a prior distribution in a Bayesian model may be sufficient to explain the origins of cross-linguistic regularities in systems of color categories.

Bayesian approaches to color category learning can be used to explore questions about how children might learn how their language partitions the space of colors and why regularities are seen in systems of color categories across languages. However, this research is still in its early stages, with many important questions remaining open. One fundamental question is how well Bayesian models can capture the generalizations that real human children make when learning color categories. Another is whether it is possible to characterize the human prior distribution over systems of color categories precisely and whether the structure of that distribution can capture cross-linguistic variation.

Cross-References

- ▶ [Berlin & Kay theory](#)
- ▶ [Color Categorical Perception](#)
- ▶ [Color Category Learning in Naming-game Simulations Comparative Color Categories World Color Survey](#)

References

1. Kay, P., Maffi, L.: Color appearance and the emergence and evolution of basic color lexicons. *Am. Anthropol.* **101**, 743–760 (1999)
2. Kersten, D., Mamassian, P., Yuille, A.: Object perception as Bayesian inference. *Annu. Rev. Psychol.* **55**, 271–304 (2004)
3. Tenenbaum, J.B., Kemp, C., Griffiths, T.L., Goodman, N.: How to grow a mind: statistics, structure, and abstraction. *Science* **331**, 1279–1285 (2011)
4. Tenenbaum, J.B., Griffiths, T.L.: Generalization, similarity, and Bayesian inference. *Behav. Brain Sci.* **24**, 629–641 (2001)
5. Xu, F., Tenenbaum, J.B.: Word learning as Bayesian inference. *Psychol. Rev.* **114**, 245–272 (2007)
6. Dowman, M.: Explaining color term typology with an evolutionary model. *Cognit. Sci.* **31**, 99–132 (2007)
7. Jones, C.D., Osorio, D., Baddeley, R.J.: Colour categorization by domestic chicks. *Proc. R. Soc. Lond. B* **268**, 2077–2084 (2001)

8. Kirby, S.: Spontaneous evolution of linguistic structure: an iterated learning model of the emergence of regularity and irregularity. *IEEE Trans. Evol. Comput.* **5**, 102–110 (2001)
9. Griffiths, T.L., Kalish, M.L.: Language evolution by iterated learning with Bayesian agents. *Cognit. Sci.* **31**, 441–480 (2007)
10. Xu, J., Griffiths, T.L., Dowman, M.: Replicating color term universals through human iterated learning. In: Ohlsson, S., Catrambone, R. (eds.) *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*, pp. 352–357. Cognitive Science Society, Austin (2010)