LEARNING TO LEARN FROM SIMILAR OTHERS: APPROXIMATE BAYESIAN COMPUTATION THROUGH BABBLING

BILL THOMPSON, HEIKKI RASILO

Artificial Intelligence Laboratory, Vrije Universiteit Brussel, Brussels, Belgium
bill@ai.vub.ac.be
heikki.rasilo@aalto.fi

To emerge and persist over cultural transmission, complex linguistic structures must be learnable from noisy, incomplete linguistic data. Rational statistical inference provides a compelling solution to inductive problems at many levels of linguistic structure. However, from an evolutionary perspective, a core question concerns how language learners could be equipped to approximate rational inferences under limited cognitive resources. We present a computational account of how *self-simulation* of linguistic data can aid otherwise challenging probabilistic inferences during language acquisition – a shortcut made possible because humans acquire language from linguistic data produced by other humans. Through an analogy with a class of computational techniques known as Approximate Bayesian Computation, we show how the capacity to *produce* language data can leverage computationally-cheap inductive leaps that approximate rational inference when *learning* language. We derive an approximate inference model for an idealised problem in the acquisition of speech sounds through babbling – a problem in which the gestural source of those sounds is invisible to the learner but must be reverse-engineered. We simulate the dynamics of cultural transmission under this model, and discuss implications for the evolution of speech and language.

1. Introduction

Language acquisition involves inductive inference under uncertainty: at many levels of language structure, the learner must reverse-engineer a richly structured linguistic model from noisy, incomplete linguistic data. Human learners – but no other learning system – reliably solve this inductive problem during early life. How this is possible remains a core question underpinning inquiry into the evolution of language. Learnability arguments underpin many theories concerning the origins of linguistic structure, from those that appeal to specialised innate *biological* evolutionary innovations (Pinker & Bloom, 1990), to those that cite the structure-forming properties of *cultural* evolution (Kirby, Griffiths, & Smith, 2014); in order to understand how linguistic structures emerge and persist among populations, we must understand the computational principles that underpin acquisition of language from culturally transmitted linguistic data.

A key recent development, with considerable implications for the evolution of language, is the discovery that domain-independent principles of rational statistical inference appear to provide solutions to many formidable inductive problems in the acquisition of structured linguistic knowledge (e.g. Perfors, Tenenbaum, & Regier, 2011). However, when guided by an evolutionary perspective, a core question arises: how might learners be equipped to implement or approximate the often demanding computations that underpin rational statistical inference? In this paper we argue that, from an inferential perspective, the language learner has a powerful trick up her sleeve: the ability to *self simulate* linguistic data. We present a computational account for how self-simulation of linguistic data can leverage approximate inductive inferences that would otherwise be intractable. The computational solution arises because language is a cultural behaviour, learned from data produced by similar others (Chater & Christiansen, 2010). We draw an analogy with a class of inferential techniques known as Approximate Bayesian Computation (ABC), which solve an analogous problem through analogous means. We focus on a concrete example – acquisition of speech sounds through babbling – and discuss implications for the evolution of speech and language.

2. Language via Rational Inference

Several aspects of linguistic structure traditionally thought to imply language-specific acquisition procedures have been shown to be learnable via rational inference. For example, Perfors et al. (2011) show that, given the statistical properties of linguistic data, an ideal rational learner exposed to these data could make the inductive leap to hierarchical phrase-structure in syntax. Likewise, statistical inference can recover the structure of phonetic categories (Vallabha, McClelland, Pons, Werker, & Amano, 2007), word segmentation boundaries (Goldwater, Griffiths, & Johnson, 2009), and structure in verb classes (Perfors, Tenenbaum, & Wonnacott, 2010), to cite just a few examples.

2.1. *An Ideal Rational Learner*

Rational inference follows an appealingly simple recipe: when evaluating a hypothesis h as an explanation for observed data d, the learner should update her *prior* belief in that hypothesis (independent of the observed data) $p(h)$ by accounting for the *likelihood* of those data under the hypothesis, $p(d|h)$, in accordance with Bayes' rule:

$$p(h|d) = p(d|h)p(h)/p(d) \ , \tag{1}$$

where $p(d)$ reflects the likelihood of the data marginalised over all possible hypotheses. This scheme for combining subjective and empirical information aligns with human behaviour in many domains. However, in many real-world learning problems, the computational cost of precisely computing the informational ingredients for this formula – $p(h)$, $p(d|h)$, and $p(d)$, – can be prohibitively demanding, for computer scientists and for human learners. As such, there is considerable interest in understanding how learners might approximate the computations that underpin rational inference under limited cognitive resources.

2.2. *Approximating Rational Inferences*

A growing body of research aims to describe a class of *rational process models* (see e.g. Griffiths, Lieder, & Goodman, 2015): these models demonstrate how simple algorithms can approximate rational inferences in inductive problems such as causal learning (Bonawitz, Denison, Gopnik, & Griffiths, 2014) and category formation (Sanborn, Griffiths, & Navarro, 2010). Likewise, there have been efforts to demonstrate how classic psychological models, such as exemplar models (Shi, Griffiths, Feldman, & Sanborn, 2010) and neural networks (Abbot, Hamrick, & Griffiths, 2013) can approximate the computations implied by Bayesian models. Several of these discoveries have drawn on analogies between human learning and inference algorithms developed in computer science and statistics. Our proposal for language follows in this vein.

3. Learning by Producing: Approximate Inference via Simulation

A core requirement for rational inference is *likelihoods* knowledge: when reverse engineering the underlying source of our observations, we must evaluate the *likelihood* of those observations under any hypothesis we entertain about their cause. However, in many realistic cases, our generative model for the data is so complex or high-dimensional that we cannot say for certain exactly how likely are our observations under the model: in formal terms, the likelihood distribution – $p(d|h)$ in equation 1 above – over possible observations cannot be efficiently computed. This is a common problem that has hindered model-fitting in a number of disciplines, such as population genetics, systems biology, and economics, in which hypothesised models (e.g. population demographic history) and their resulting data (e.g. gene-sequence data) are high-dimensional, and the data likelihoods intractable. We suggest that – given the complex structures that underpin language, and the open-endedness of linguistic data – the language learner would often find herself in a similar position. While ideal learner models generally have access to these computations, how is the language learner to fare if she does not? The solution we explore here rests on the fact that, in the case of language, the *learner* can also be a *producer*, able to test hypotheses by self-simulation of linguistic data.

3.1. *Computation in the Face of Unknown Likelihoods*

Advances in computer science and statistics have uncovered methods for performing inference in the case of unknown likelihood quantities. In particular, considerable progress has been made for a special class of cases: those in which, though we may *not* be able to compute the likelihood of our observations exactly, we *are* nevertheless in a position to *simulate data from the model*. Various simple inferential procedures have been developed to perform approximate inference on the basis of data simulation, generally known as Approximate Bayesian Computation (see e.g. Beaumont, 2010, for a review of these methods). We propose that aspects of language acquisition fall into this class of problem. Language is culturally transmitted: it is acquired through inference over data generated by other humans. Since teacher and learner share anatomical and cognitive properties, the learner is in a position to simulate linguistic data under a mechanism that approximates the generating mechanism, and could leverage this capacity during inference. ABC methods provide a computational framework to demonstrate *how*, and a plausible model for several inductive problems in language.

3.2. *A Simple Rejection-ABC Algorithm*

Here we describe a simple ABC algorithm based on rejection sampling: in order to demonstrate the principles behind our proposal, we aim for maximal simplicity throughout the examples in this paper, and discuss more sophisticated extensions in section 5. Given an observed dataset $\mathbf{y} = \{y_1, \dots y_n\}$ generated by a mechanism with an unknown parameter θ, the inductive problem is to characterise the posterior distribution $p(\theta|\mathbf{y}) = p(\mathbf{y}|\theta)p(\theta)/p(\mathbf{y})$: here $p(\theta)$ is the prior distribution over model parameters θ, $p(\mathbf{y}|\theta)$ is the likelihood function for the data conditional on the parameter θ, and $p(\mathbf{y})$ is a normalising constant. Assuming it is impractical to evaluate the likelihood function $p(\mathbf{y}|\theta)$ directly, but simple to *generate* data $\mathbf{x}$ from the distribution $p(\mathbf{x}|\theta)$, then it is possible to collect samples from the posterior distribution by repeating the following procedure:

1. Generate $\theta_i \sim p(\theta)$
2. Generate $\mathbf{x_i} \sim p(\mathbf{x_i}|\theta_i)$
3. Store θ_i if $\rho(S(\mathbf{x_i}), S(\mathbf{y})) < \epsilon$.

Here $S(\cdot)$ represents summary statistics of the dataset, $\rho(\cdot, \cdot)$ is a difference measure between the simulated and observed datasets, and ϵ is an acceptance tolerance level. The stored values of θ then represents a set of samples that approximates the posterior distribution $p(\theta|\mathbf{y})$. This process is akin to one of trial-and-error, and can be summarised intuitively as follows: keep guessing parameter values, producing data, and storing any guesses that led to data which is close to the original observation. In the next section, we introduce our illustrative inductive problem – speech sound acquisition through babbling – and show how this algorithm can be understood as a model of learning that approximates rational inference in that context.

4. Speech Sound Acquisition through Babbling

One of the first inductive leaps the language learner must make is to learn a structured system of speech sounds. For example, on the basis of the speech sounds it encounters, the child must reverse engineer arrangements of the articulatory apparatus required to produce the system of sounds used in its language.

In traditional terminology, the learner must acquire the *backward mapping* from sounds to articulators. To do this the learner must know something of the *forward mapping* from articulators to sounds - corresponding the likelihood distribution in the Bayesian framework. Babbling is typically thought to allow the child to learn the relationship between articulatory gestures and their acoustic counterparts (Vihman, 1991). Several computational models have been proposed for learning the mappings in a babbling phase, such as the neural network models of Guenther (2006) and Kröger, Kannampuzha, and Neuschaefer-Rube (2009), and these learned mappings are later used to invert observed native language phonemes back to articulation, enabling imitation. Our framework offers a bridge between these implementation-level models and computational-level rational analyses.

4.1. *Noise in Production of a Target Speech Sound*

Here we describe an idealised model of speech perception which qualitatively captures the nature of the inductive problem: inference of an underlying unknown quantity from a dataset of noisy exemplars. Our strategy is to study a simple, abstract model of inference in which we, the modelers, know the likelihood distribution, but assume the learner does not and must perform inference by data simulation as described above. Assume the speaker aims to produce a target speech sound τ: τ reflects a single phonetic feature on a continuous range, such as voice-onset time or an absolute formant value. Realisations of this speech sound are noisy: when the speaker aims for τ, it produces a sound that is normally distributed with mean τ and variance σ^2. The learner must reverse engineer the underlying intended target τ from a set of n exemplars $\mathbf{y} = \{y_1, y_2, \ldots, y_n\}$.

4.2. *Approximate Bayesian Computation through Babbling*

Given $\mathbf{y}$, an ideal rational learner estimating τ would compute the posterior distribution $p(\tau|\mathbf{y}) \propto p(\mathbf{y}|\tau)p(\tau)$. Assume a Gaussian prior - $p(\tau) \propto \mathcal{N}(\tau_0, \sigma_0^2)$ – over the range of *possible* values for τ. The prior may represent any constraints (e.g. anatomical or cognitive constraints) that influence speech production. Likewise allow a likelihood function defined by normally distributed noise around the target – $p(\mathbf{y}|\tau) \propto \mathcal{N}(\bar{y}; \tau, \sigma^2/n)$. Our hypothetical language learner does not know the likelihood function, but can sample from its own data-producing mechanism. Following the procedure outlined in section 3.2, the babbling learner approximates rational inferences as follows:

1. Sample hypothesised target from prior $\hat{\tau}_k \sim \mathcal{N}(\tau_0, \sigma_0^2)$
2. Babble m speech sounds $\mathbf{x} = \{x_1, \ldots, x_m\} \sim \mathcal{N}(\mathbf{x}; \hat{\tau}_k, \sigma_{self}^2)$
3. Store hypothesised $\hat{\tau}_k$ if $|\bar{x} - \bar{y}| < \epsilon$.

This process results in a set of samples that approximates the posterior distribution $p(\tau|\mathbf{y})$. Figure 1 presents a simple visual example. The accuracy of this approximation depends on a number of factors, including the number of iterations (N), the number of babbled sounds (step 2 - m), the acceptance criteria and tolerance rate (step 3 - ϵ), and how closely the learner's data simulation mechanism ($p(\hat{\tau}_k), \sigma^2_{self}$) matches that which produced the observed data (τ, σ^2).

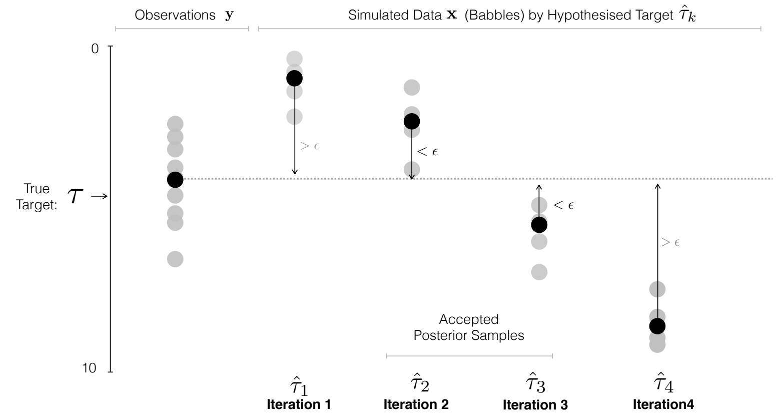

Figure 1. A schematic example of the ABC algorithm as a model of learning: circles represent datapoints – the observed speech sound (leftmost column), and the simulated speech sounds (columns 2-5) – with the sample mean given by the black circle. Only the second and third hypothesised targets τ_2 & τ_3 are accepted as posterior samples, since the mean of the resulting babbles is close to the observed data mean (dotted line).

4.3. *Cultural Transmission Among a Population*

How does the approximation procedure affect the linguistic structures that emerge over cultural transmission? Existing results for cultural transmission under ideal rational inference (Griffiths & Kalish, 2007) show that the population level distribution of linguistic structures should converge toward the learners' prior distribution: the population level distribution of sound estimates should reflect the prior over τ. By substituting arbitrary parameter values $\tau_0 = 2$ and variance $\sigma_0^2 = 1$ into the Gaussian prior, we can ask how the approximate inference procedure influences this relationship between individuals and populations. Figure 2 shows the mean and variance in the population-level distribution of τ, averaged over 50 simulations of cultural transmission at each point, as a function of the acceptance threshold ϵ. In each of these simulations, 500 generations of single agents in an iterated learning chain (Kirby et al., 2014) transmit τ by listening to the sounds produced by the previous learner, drawing a sample $\hat{\tau}$ from the set of accepted ABC posterior samples, and producing more linguistic data conditional on $\hat{\tau}$ from which the next learner learns. Given the symmetry in the Gaussian representations, the population mean for τ always approximates the prior mean. However,

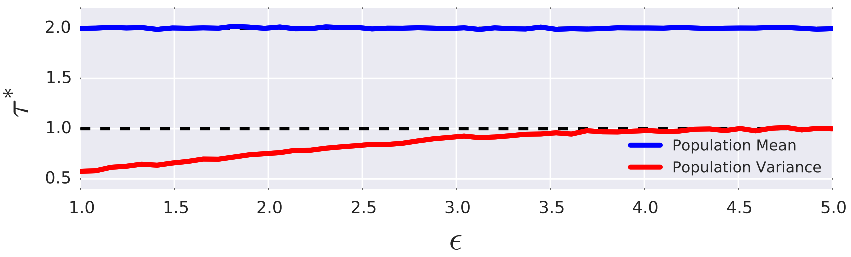

Figure 2. Mean (blue) and variance (red) for τ^*, the population-level distribution of τ, over simulations of 500 generations of learners, average over 50 replications, as a function of ϵ. Dashed lines give the shared prior mean and variance. $n = 20, N = 200, m = 5, \sigma^2_{self} = 0.1$. See section 4.2 for details of these parameters.

the variance in the population-level distribution of τ only converges to the prior variance under more tolerant acceptance thresholds (higher ϵ). In intuitive terms: if only estimates very close to the target sound are accepted, little cultural evolution occurs; if the tolerance level is weaker, the speech of learners drifts over time to reflect the prior distribution (speech sounds that are easier to produce).

5. Conclusion and Future Directions

We presented a psychologically lightweight computational account for how learners can converge on approximate rational inferences via self-simulation of linguistic data. The fact that learner is also producer broadens the class of learnable linguistic structures that can emerge, persist, and be shaped over cultural transmission: for example, computational principles such as these may make structured systems of speech sounds learnable, and thus evolvable, in the face of unobservable speech-production processes.

The account we developed is maximally abstract: in future work we hope to apply these ideas to specific inductive problems, using articulatory models and acoustic data. Likewise, the ABC procedure we considered is the simplest possible: a fruitful continuation of this research might explore ABC methods which include reinforcement learning, which is known to play an important role during infant speech acquisition (Goldstein, King, & West, 2003), for example. Likewise, the framework provides a natural basis to capture and quantify the effects of *differences* between data generation and simulation mechanisms. For instance, differences between infant and care giver vocal tract morphologies may hinder direct comparison of the spectral features of babbles and adult phonemic categories: reinforcement by caregivers, who know the correspondence between the their own voices and the infant's, may help in category formation, and could be modeled in this framework. A key advantage of our computational proposal is that, when applied to specific problems, existing results for ABC methods make specific predictions about the trajectory and outcome of learning. Our proposal also predicts, for example, that the order in which articulators develop may influence which speech sounds are acquired first (Esling, Benner, & Moisik, 2015). More generally, the computational principles we have described could apply to any human (or non-human, e.g. birdsong) behaviour that is culturally transmitted via a production mechanism that the learner cannot observe but can simulate herself. Understanding the computational principles that allow data-simulation - in speech or via more general capacities such as pragmatic reasoning – to aid learning from others is an important step toward understanding the evolution of our flagship cultural behaviour, human language.

References

Abbot, J. T., Hamrick, J. B., & Griffiths, T. L. (2013). Approximating Bayesian inference with a sparse distributed memory system. In M. Knauf, M. Pauen, M. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th annual conference of the cognitive science society.* Berlin: Cognitive Science Society.

Beaumont, M. A. (2010). Approximate Bayesian Computation in Evolution and Ecology. *Annual Review of Ecology, Evolution, and Systematics*, *41*(1), 379–406.

Bonawitz, E., Denison, S., Gopnik, A., & Griffiths, T. L. (2014). Win-Stay, Lose-Sample: a simple sequential algorithm for approximating Bayesian inference. *Cognitive psychology*, *74*, 35–65.

Chater, N., & Christiansen, M. H. (2010). Language acquisition meets language evolution. *Cognitive science*, *34*(7), 1131–57.

Esling, J. H., Benner, A., & Moisik, S. R. (2015). Laryngeal Articulatory Function and Speech Origins. In H. Little (Ed.), *Proceedings of the 18th international congress of phonetic sciences satelite event: The evolution of phonetic capacities.* Glasgow, UK.

Goldstein, M. H., King, A. P., & West, M. J. (2003). Social interaction shapes babbling: testing parallels between birdsong and speech. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(13), 8030–5.

Goldwater, S., Griffiths, T. L., & Johnson, M. (2009). A Bayesian framework for word segmentation: exploring the effects of context. *Cognition*, *112*(1), 21–54.

Griffiths, T. L., & Kalish, M. L. (2007). Language evolution by iterated learning with bayesian agents. *Cognitive science*, *31*(3), 441–80.

Griffiths, T. L., Lieder, F., & Goodman, N. D. (2015). Rational use of cognitive resources: levels of analysis between the computational and the algorithmic. *Topics in cognitive science*, *7*(2), 217–29.

Guenther, F. H. (2006). Cortical interactions underlying the production of speech sounds. *Journal of communication disorders*, *39*(5), 350–65.

Kirby, S., Griffiths, T., & Smith, K. (2014). Iterated learning and the evolution of language. *Current opinion in neurobiology*, *28C*, 108–114.

Kröger, B. J., Kannampuzha, J., & Neuschaefer-Rube, C. (2009). Towards a neurocomputational model of speech production and perception. *Speech Communication*, *51*(9), 793–809.

Perfors, A., Tenenbaum, J. B., & Regier, T. (2011). The learnability of abstract syntactic principles. *Cognition*, *118*(3), 306–38.

Perfors, A., Tenenbaum, J. B., & Wonnacott, E. (2010). Variability, negative evidence, and the acquisition of verb argument constructions. *Journal of Child Language*, *37*(03), 607–642.

Pinker, S., & Bloom, P. (1990). Natural language and natural selection. *Behavioral and Brain Sciences*, *13*(4), 707–784.

Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2010). Rational approximations to rational models: alternative algorithms for category learning. *Psychological review*, *117*(4), 1144–67.

Shi, L., Griffiths, T. L., Feldman, N. H., & Sanborn, A. N. (2010). Exemplar models as a mechanism for performing Bayesian inference. *Psychonomic bulletin & review*, *17*(4), 443–64.

Vallabha, G. K., McClelland, J. L., Pons, F., Werker, J. F., & Amano, S. (2007). Unsupervised learning of vowel categories from infant-directed speech. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(33), 13273–8.

Vihman, M. (1991). Ontogeny of Phonetic Gestures: Speech Production. In N. Hillsdale (Ed.), *Modularity and the model theory of speech perception: Proceedings of a conference to honor alvin m. liberman* (pp. 69–84). Erlbaum.