Partial Operator Induction with Beta Distributions

Nil Geisweiller[1,2,3]

[1] SingularityNET Foundation
[2] OpenCog Foundation
[3] Novamente LLC

Abstract. A specialization of Solomonoff Operator Induction, driven by OpenCog model representation but expected to be more broadly useful, considering partial operators described by Beta distributions is introduced. The problem of taking into account partial operators in the prediction estimate is presented. This problem turns out to be non-trivial. A simplistic solution with a heuristic to estimate the Kolomogorov complexity of completions of partial models is given.

Keywords: Solomonoff Operator Induction · Beta Distribution · Bayesian Averaging.

1 Introduction

Rarely natural intelligent agents attempt to construct complete models of their environment. Often time they compartmentalize their knowledge into contextual rules and make use of them without worrying about the details of the assumingly remote and irrelevant parts of the world.

This is typically how AGI Prime, aka OpenCog Prime, the AGI agent implemented over the OpenCog framework may utilize knowledge [3]. The models we are specifically targeting here are conditional probabilities, or to be more precise probability distributions over conditional probabilities, or *second order* conditional probabilities. Maintaining second order probabilities is how OpenCog accounts for uncertainties [7] and by that properly manages weighting knowledge from heterogeneous sources, balancing exploitation and exploration and so on.

We will sometimes call these models, rules, understanding that they actually represent second order conditional probabilities. Here are some examples of rules

1. If the sun shines, then the temperature rises
2. If the sun shines and there is no wind, then the temperature rises
3. If the sun shines and the agent is in a cave, then the temperature rises

These 3 rules have different degrees of truth. The first one is often true, the second is nearly always true and the last one is rarely true. The traditional way to quantify these degrees of truth is to assign probabilities. In practice though these probabilities are unknown, and instead one may only assign probability estimates based on limited evidence. Or, according to the OpenCog design, one may assign distributions over probabilities, capturing their degree of certainty. The wider the less certain, the narrower the more certain.

Once degrees of truth are properly represented, an agent should be able to utilize these rules to predict and operate in its environment. This raises a question. How to choose between rules? Someone wanting to predict whether the temperature will rise will have to make a choice. If one is in a cave, should he/she follow the third rule? Why not the first one which is valid, or assuming there is no wind, maybe the second?

Systematically picking the rule with the narrowest context (like being in a cave) is not always right. Indeed, the narrower the context the less evidence we have, the broader the uncertainty, the more prone to overfitting such rule might be.

1.1 Contribution

In this paper we attempt to address this issue by adapting Solomonoff Operator Induction [8] for a special class of operators representing such rules. These operators have two particularities. First, their outcomes are second order probabilities, specifically Beta distributions. Second, they are partial, that is they are only defined over a subset of observations, the available observations meeting the conditions of a given rule. For instance if the goal is to predict the consequences of some actions taken in the context of riding bicycle. Rules capturing that context and no broader will not be able to account for observations made in excluded contexts, such as walking. This latter particularity turns out to be very difficult to address, and the solution we offer is very lacking but presented nevertheless as a start.

1.2 Overview

In Section 2 we briefly recall Solomonoff Operator Induction, Beta distributions. In Section 3 we introduce our specialization of Solomonoff Operator Induction for partial operators with Beta distributions. Finally in Section 4 we conclude and present some directions for further research.

2 Recall

2.1 Solomonoff Operator Induction

Solomonoff Universal Operator Induction [8] is a general, parameter free induction method that has been shown to theoretically converge to any true computable distribution. It is a special case of Bayesian Model Averaging [5] though is universal in the sense that the models across which the averaging is taking place are Turing-complete.

Let us recall its formulation, using the same notations as in the original paper of Solomonoff (Section 3.2 of [8]). Given a sequence of n questions and answers $(Q_i, A_i)_{i\in[1,n]}$, and a countable family of operators O^j (the superscript j denotes the j^{th} operator, not the exponentiation) computing partial functions mapping pairs of question and answer to probabilities, then one may estimate the probability of the next answer A_{n+1} given new question Q_{n+1} as follows

$$\hat{P}(A_{n+1}|Q_{n+1}) = \sum_j a_0^j \prod_{i=1}^{n+1} O^j(A_i|Q_i) \tag{1}$$

where a_0^j is the prior of the j^{th} operator (its probability after zero observation). Using Hutter's convergence theorems to arbitrary alphabets [6] it can be shown that such estimate rapidly converges to the true probability.

Let us rewrite this equation by making the prediction term and the likelihood explicit

$$\hat{P}(A_{n+1}|Q_{n+1}) = \sum_j a_0^j l^j O^j(A_{n+1}|Q_{n+1}) \tag{2}$$

where $l^j = \prod_{i=1}^{n} O^j(A_i|Q_i)$ is the likelihood, the probability of the data given the j^{th} operator.

Remark 1. In the remaining of the paper the superscript j is always used to denote the index of the j^{th} operator. Sometimes, though in a consistent manner, it is used as subscript. All other superscript notations not using j denote exponentiation.

2.2 Beta Distribution

Beta distributions [1] are convenient to model probability distributions over probabilities, i.e. second order probabilities. In particular, given a prior over a probability p of some event, like a coin toss to head, defined by a Beta distribution, and a sequence of experiments, like n coin tosses, the posterior of p is still a Beta distribution. For that reason the Beta distribution is called a *conjugate prior* for the binomial distribution.

Let us recall the probability density and cumulative distribution functions of the Beta distribution as it will be useful later on.

Prior and Posterior Probability Density Function The probability density function (pdf) of the Beta distribution with parameters α and β, is

$$f(x;\alpha,\beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{\mathrm{B}(\alpha,\beta)} \tag{3}$$

where x is a probability and $\mathrm{B}(\alpha,\beta)$ is the beta function

$$\mathrm{B}(\alpha,\beta) = \int_0^1 p^{\alpha-1}(1-p)^{\beta-1}dp \tag{4}$$

One may see that multiplying the density by the likelihood

$$x^m(1-x)^{n-m} \tag{5}$$

of a particular sequence of n experiments with m positive outcomes, is also a Beta distribution

$$f(x; m+\alpha, n-m+\beta) \propto x^{m+\alpha-1}(1-x)^{n-m+\beta-1} \tag{6}$$

Cumulative Distribution Function The cumulative distribution function (cdf) of the Beta distribution is

$$I_x(\alpha, \beta) = \frac{\mathrm{B}(x; \alpha, \beta)}{\mathrm{B}(\alpha, \beta)} \tag{7}$$

where $\mathrm{B}(x; \alpha, \beta)$ is the incomplete beta function

$$\mathrm{B}(x; \alpha, \beta) = \int_0^x p^{\alpha-1}(1-p)^{\beta-1}dp \tag{8}$$

I_x is also called the regularized incomplete beta function.

3 Partial Operator Induction with Beta Distributions

In this section we introduce a specialization of Solomonoff Operator Induction for partial operators describing second order distributions.

3.1 Second Order Probability Estimate

Let us first modify the Solomonoff Operator Induction probability estimate to become a second order probability estimate. This is crucial to maintain the uncertainty surrounding that estimate. It directly follows from Eq. 2 of Section 2.1, that the cumulative distribution function of the probability estimate of observing answer A_{n+1} given question Q_{n+1} is

$$\hat{cdf}(A_{n+1}|Q_{n+1})(x) = \sum_{O^j(A_{n+1}|Q_{n+1}) \leq x} a_0^j l^j \tag{9}$$

Due to O^j not being complete in general $\hat{cdf}(A_{n+1}|Q_{n+1})(1)$ may not be equal to 1. It means that some normalization will need to take place in practice. That is even more true in our case since, as will be shown further below, the operators taken into consideration are restricted to a subclass. Also, obviously the continuity or the differentiability of $\hat{cdf}(A_{n+1}|Q_{n+1})$ do not generally hold. What matters is that a spread of probabilities is represented to properly account for the uncertainty of that estimate. It is expected that the breadth would be wide at first, and progressively shrinks, fluctuating depending on the novelty of the contexts, as measure as more questions and answers get collected.

3.2 Continuous Parameterized Operators

Let us now extend this for parameterized operators, so that each operator is a second order distribution. Let us consider a subclass of parameterized operators such that, if p is the parameter of operator O^j_p, the result of the conditional probability of A_{n+1} given Q_{n+1} is

$$O^j_p(A_{n+1}|Q_{n+1}) = p \tag{10}$$

We do that to later consider Beta distribution operators. The reason for this assumption will become clearer in Section 3.3. Given that assumption, the cumulative distribution function of the estimate $\hat{cdf}(A_{n+1}|Q_{n+1})$ becomes

$$\hat{cdf}(A_{n+1}|Q_{n+1})(x) = \sum_j a^j_0 \int_0^x f_p l^j_p dp \tag{11}$$

where f_p is the prior density of p, and $l^j_p = \prod_{i=1}^n O^j_p(A_i|Q_i)$ is the likelihood of the data according to the j^{th} operator with parameter p.

Proof. Consider continuous families of parameterized operators combined with Eq. 10. Let us start with the discrete case

$$\hat{cdf}(A_{n+1}|Q_{n+1})(x) = \sum_{O^j_p(A_{n+1}|Q_{n+1}) \leq x} a^j_0 f_p l^j_p \Delta p \tag{12}$$

where the sum runs over all j and p by steps of Δp such that $O^j_p(A_{n+1}|Q_{n+1}) \leq x$. Assuming that a^j_0 does not depends on p, it can be moved in its own sum

$$\hat{cdf}(A_{n+1}|Q_{n+1})(x) = \sum_j a^j_0 \sum_{O^j_p(A_{n+1}|Q_{n+1}) \leq x} f_p l^j_p \Delta p \tag{13}$$

now the second sum only runs over p. Due to Eq. 10 this can be simplified into

$$\hat{cdf}(A_{n+1}|Q_{n+1})(x) = \sum_j a^j_0 \sum_{p \leq x} f_p l^j_p \Delta p \tag{14}$$

which is turns into Eq. 11 when Δp tends to 0.

Using continuous integration may seem like a departure from Solomonoff Induction. First, it does not correspond to a countable class of models. And second, the Kolmogorov complexity of p, that would in principle determine its prior, is likely chaotic and very different than how priors are typically defined over continuous parameters in Bayesian inference. In practice however integration is discretized and values are truncated up to some fixed precision. Moreover any prior can probably be made compatible with Solomonoff induction by selecting an adequate Turing machine of reference.

3.3 Operators as Beta Distributions

We have now all we need to fit our rules, second order conditional probabilities, into Solomonoff Operator Induction.

First, we need to assume that operators are partial, that is the j^{th} operator is only defined for a subset of n^j questions, those that meet the conditions of the rule. For instance, with the rule

– If the sun shines, then the temperature rises

questions and answers pertaining to what happens at night will be ignored.

Second, we assume that operators only represent whether answer A_i is equal to A_{n+1} or not. In reality, OpenCog rules manipulate predicates (generally fuzzy predicates but that is let aside), and the questions asked are: given any instance holding some property P, what are the odds that it holds property Q. So we may assume that the property of interest Q corresponds to whether $A_{n+1} = A_i$, or equivalently that answers are Boolean. This allows an operator to be modeled as a Beta distribution, with cumulative distribution function

$$cdf_{O^j} = I_x(m^j + \alpha, n^j - m^j + \beta) \tag{15}$$

where m^j is the number of times $A_{n+1} = A_i$ for the subset of n^j questions defined for the j^{th} operator, α and β are the parameters of the Beta distribution prior. This corresponds to the definition of OpenCog truth values (see Chapter 4 of the PLN book [4]).

3.4 Handling Partial Operators

When attempting to use such operators we still need to account for their partiality. Although Solomonoff Operator Induction does in principle encompass partial operators[4], it does so insufficient, in our case anyway. Indeed, if a given operator cannot compute the conditional probability of some answer question pair, the contribution of that operator may simply be ignored in the estimate. This does not work for us since partial operators (rules over restricted contexts) might carry significant predictive power and should not go to waste.

To the best of our knowledge, the existing literature does not cover that problem. The Bayesian inference literature contains in-depth treatments about how to properly consider missing data [10]. Unfortunately, they do not directly apply here because our assumptions are different. In particular, here, data omission depends on the model. However, general principles such as modeling missing data and taking into account these models in the inference process, can be applied. Let us attempt to do that in a by explicitly representing the portion of the likelihood over the missing data according to the j^{th} operator by a term, denoted r^j. In the rest of the paper rather than calling these data *missing* we prefer to

[4] more by necessity, since the set of partial operators are countable, while the set of complete ones are not

denominate them as *unexplained* or *unaccounted* by operator O^j, which better captures our assumption. Let us replace the likelihood in Eq. 11 by

$$l_p^j = p^{m^j}(1-p)^{n^j-m^j} r^j \tag{16}$$

where the binomial term account for the likelihood of the explained observations by the j^{th} operator with parameter p, and r^j accounts for the likelihood of the unexplained observations. One may notice that r^j does not depends on p. Such assumption tremendously simplifies the analysis and is somewhat reasonable to make. It means that the *completion* of the model is independent on its pre-existing part. By completion we mean computation that explains the unaccounted data. For instance a possible completion of the rule

- If the sun rises, then the temperature rises

could be

- Otherwise, the temperature falls

Using this likelihood the cumulative distribution function of the estimate of A_{n+1} knowing Q_{n+1} becomes

$$\hat{cdf}(A_{n+1}|Q_{n+1})(x) = \sum_j a_0^j \int_0^x f_p p^{m^j}(1-p)^{n^j-m^j} r^j dp \tag{17}$$

Choosing a Beta distribution as the prior of f_p simplifies the equation as the posterior remains a Beta distribution

$$f_p = f(p;\alpha,\beta) \tag{18}$$

where f is the pdf of the Beta distribution as define in Eq. 3. Usual priors are Bayes' with $\alpha = 1$ and $\beta = 1$, Haldane's with $\alpha = 0$ and $\beta = 0$ and Jeffreys' with $\alpha = \frac{1}{2}$ and $\beta = \frac{1}{2}$. The latter is the most accepted due to being *uninformative* in some sense [9]. We do not need to commit to a particular one at that point and let the parameters α and β free, giving us

$$\hat{cdf}(A_{n+1}|Q_{n+1})(x) = \sum_j a_0^j \int_0^x \frac{p^{\alpha-1}(1-p)^{\beta-1}}{\mathrm{B}(\alpha,\beta)} p^{m^j}(1-p)^{n^j-m^j} r^j dp \tag{19}$$

r^j can be moved out of the integral and the constant $\mathrm{B}(\alpha,\beta)$ can be ignored on the ground that our estimate requires some normalization anyway

$$\hat{cdf}(A_{n+1}|Q_{n+1})(x) \propto \sum_j a_0^j r^j \int_0^x p^{m^j+\alpha-1}(1-p)^{n^j-m^j+\beta-1} dp \tag{20}$$

$\int_0^x p^{m^j+\alpha-1}(1-p)^{n^j-m^j+\beta-1} dp$ is the incomplete Beta function with parameters $m^j+\alpha$ and $n^j-m^j+\beta$, thus

$$\hat{cdf}(A_{n+1}|Q_{n+1})(x) \propto \sum_j a_0^j r^j B(x; m^j+\alpha, n^j-m^j+\beta) \tag{21}$$

Using the regularized incomplete beta function we obtain

$$\hat{cdf}(A_{n+1}|Q_{n+1})(x) \propto \sum_j a_0^j r^j I_x(m^j+\alpha, n^j-m^j+\beta)B(m^j+\alpha, n^j-m^j+\beta) \quad (22)$$

As I_x is the cumulative distribution function of O^j (Eq. 15), we finally get

$$\hat{cdf}(A_{n+1}|Q_{n+1})(x) \propto \sum_j a_0^j r^j cdf_{O^j}(x)B(m^j+\alpha, n^j-m^j+\beta) \quad (23)$$

We have expressed our cumulative distribution function estimate as an averaging of the cumulative distribution functions of the operators. This averaging is hopefully close to optimal (since the operators are a subclass of Turing-complete operators optimality cannot be guarantied), and most importantly it captures the uncertainty of the estimate.

We still need to address r^j, the likelihood of the unaccounted data. In theory, the right way to model r^j would be to consider all possible completions of the j^{th} operator, but that is intractable. One would be tempted to simply ignore r^j, however, as we have already observed in some preliminary experiments, this gives an unfair advantage to rules that have a lot of unexplained data, and thus make them more prone to overfitting. This is true even in spite of the fact that such rules exhibit more uncertainty, that is their second order probability have wider dispersion due to having less evidence.

3.5 Perfectly Explaining Unaccounted Data

Instead we attempt to consider the most prominent completions. For now we consider completions that perfectly explain the unaccounted data. Moreover, to simplify further, we assume that unaccounted answers are entirely determined by their corresponding questions. This is generally not true, the same question may relate to different answers. But under such assumptions r^j becomes 1. This may seem equivalent to ignoring r^j unless the complexity of the completion is taken into account. What that means is that we must consider, not only the complexity of the rule but also the complexity of the completion. Unfortunately calculating that complexity (that is the Kolmogorov complexity) of is intractable. To work around that we estimate it with a simple heuristic

$$a_0^j = K(O^j) + v_j^{(1-k)} \quad (24)$$

where $K(O^j)$ is the Kolmogorov complexity of the j^{th} operator, v_j is the size of the unaccounted data by the j^{th} operator, and k is a *compressability* parameter. If $k = 0$ then the unaccounted data are incompressible. If $k = 1$ then the unaccounted data can be compressed to a single bit. It is a very crude heuristic and is not parameter free, but it is simple and computationally lightweight. When applied to experiments (not described here due to their embryonic nature and due to space limitation) a value of $k = 0.5$ was actually shown to be satisfactory.

4 Conclusion

We have introduced a specialization of Solomonoff Operator Induction over operators with the particularities of being partial and being modeled by Beta distributions. While doing so we have uncovered an interesting problem, how to include the contributions of partial operators in the averaging. This problem appears to have no obvious solution, is manifestly under-addressed by the research community, and is yet important in practice. Although the solution we provide is very lacking (crudely estimating the Kolmogorov complexity of a perfect completion) we hope that it may motivate further research. Even though, ultimately, it is expected that this problem is hard enough that it may require some form of meta-learning [2], improvements in the heuristic by, for instance, considering completions reusing available models that do explain some unaccounted data could help.

Experiments using this estimate are currently being carried out in the context of inference control meta-learning within the OpenCog framework and will be presented in future publications.

References

1. Abourizk, S., Halpin, D., Wilson, J.: Fitting beta distributions based on sample data. Journal of Construction Engineering and Management **120** (1994)
2. Goertzel, B.: Probabilistic growth and mining of combinations: A unifying meta-algorithm for practical general intelligence. Artificial General Intelligence: 9th International Conference pp. 344–353 (2016)
3. Goertzel, B., Geisweiller, N., Monroe, E., Duncan, M., Yilma, S., Dastaw, M., Bayetta, M., Belayneh, A., Ikle, M., Yu, G.: Speculative scientific inference via synergetic combination of probabilistic logic and evolutionary pattern recognition. Artificial General Intelligence: 8th International Conference pp. 80–89 (2015)
4. Goertzel, B., Ikle, M., Goertzel, I.F., Heljakka, A.: Probabilistic Logic Networks. Springer US (2009)
5. Hoeting, J.A., Madigan, D., Raftery, A.E., Volinsky, C.T.: Bayesian model averaging: a tutorial. Statist. Sci. **14**(4), 382–417 (1999)
6. Hutter, M.: Optimality of universal bayesian sequence prediction for general loss and alphabet. Journal of Machine Learning Research **4**, 971–1000 (2003)
7. Ikle, M., Goertzel, B.: Probabilistic quantifier logic for general intelligence: An indefinite probabilities approach. Artificial General Intelligence: First International Conference pp. 188–199 (2008)
8. J. Solomonoff, R.: Three kinds of probabilistic induction: Universal distributions and convergence theorems. Comput. J. **51**, 566–570 (2008)
9. Jeffreys, H.: An Invariant Form for the Prior Probability in Estimation Problems. Proceedings of the Royal Society of London Series A **186**, 453–461 (1946)
10. Schafer, J.L., Graham, J.W.: Missing data: Our view of the state of the art. Psychological Methods p. 177 (2002)