Partial Operator Induction with Beta Distributions

Nil Geisweiller[1,2,3] [0000-0001-5041-6299]

[1] SingularityNET Foundation, The Netherlands
nil@singularitynet.io
[2] OpenCog Foundation, Delaware, USA
[3] Novamente LLC, Maryland, USA

Abstract. A specialization of Solomonoff Operator Induction considering partial operators described by second order probability distributions, and more specifically Beta distributions, is introduced. An estimate to predict the second order probability of new data, obtained by averaging the second order distributions of partial operators, is derived. The problem of managing the partiality of the operators is presented. A simplistic solution based on estimating the Kolmogorov complexity of *perfect completions* of partial operators is given.

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

1 Introduction

Rarely do 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 PrimeAGI, aka OpenCog Prime, the AGI agent implemented over the OpenCog framework may utilize knowledge [4]. The models we are specifically targeting here are rules describing *second order* conditional probabilities, probabilities over probabilities. Maintaining second order probabilities is how OpenCog accounts for uncertainties [8] and by that properly manages cognitive tasks such as integrating knowledge from heterogeneous sources, balancing exploitation and exploration and so on. 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 I am 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. Another possibility is to assign second order probabilities, distributions over probabilities as to capture their degrees of certainty. The wider the distribution the less certain, the narrower the more certain.

Once degrees of truth and confidence 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 still 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 it might be.

In this paper we attempt to address this issue by adapting Solomonoff Operator Induction [9] 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 encompassed by their associated contexts.

The remaining of the paper is organized as follows. In Section 2 we briefly recount the idea of Solomonoff Operator Induction and in Section 3, the definition and properties of Beta distributions. In Section 4 we introduce our specialization of Solomonoff Operator Induction for partial operators with Beta distributions. An estimate of the second order probability predicting new data, obtained by averaging the second order probabilities of these partial operators, is derived. Then the problem of dealing with partial operators is presented and somewhat minimally addressed. Finally, in Section 5 we conclude and present some directions for research.

2 Solomonoff Operator Induction

Solomonoff Universal Operator Induction [9] is a general, parameter free induction method shown to theoretically converge to the true distribution, the source underlying the generation of a sequence of symbols, provided that such a source is computable. It is a special case of Bayesian Model Averaging [6] though is universal in the sense that the class of models across which the averaging is taking place is Turing complete.

Let us recall its formulation, using the same notations as in the original paper of Solomonoff (Section 3.2 of [9]). 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, one may estimate the probability of the next answer A_{n+1} given 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, generally approximated by $2^{-K(O^j)}$ where K is the Kolmogorov complexity [11]. Using Hutter's convergence theorem to arbitrary alphabets [7] it can be shown that such estimate rapidly converges to the true probability.

Let us rewrite Eq. 1 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.

3 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 tossing coins, 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.

3.1 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 notice 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 with probability x, is also a Beta distribution

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

3.2 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 [13].

4 Partial Operator Induction with Beta Distributions

In this section we introduce our specialization of Solomonoff Operator Induction for partial operators describing second order distributions, and more specifically Beta distributions. An estimate of the second order conditional probability of the next data is derived, however it contains unknown terms, the likelihoods of the unaccounted data by partial operators, themselves estimated by a simplistic heuristic.

4.1 Second Order Probability Estimate

Let us first modify the Solomonoff Operator Induction probability estimate to represent a second order probability. This allows us to maintain, and ultimately propagate to efferent cognitive processes, the uncertainty of that estimate. It directly follows from Eq. 2 of Section 2, 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, that is even more true in practice since only a fraction of the operator space is typically explored. Also, we need not to worry about properties such as the continuity or the differentiability of $\hat{cdf}(A_{n+1}|Q_{n+1})$. What matters is that a spread of probabilities is represented to account for the uncertainty. It is expected that the breadth would be wide at first, and progressively shrinks, fluctuating depending on the novelty of the data, as measure as more questions and answers get collected.

4.2 Continuous Parameterized Operators

Let us now extend the definition of this estimate for parameterized operators to describe second order distributions. 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 p. Doing so will enable us to consider operators as Beta distribution later on, in Section 4.3.

Theorem 1. *Given a family of parameterized operators O^j_p such that*

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

and the prior of O^i_p is $a^j_0 f_p$ where f_p is the prior density of p, the cumulative distribution function of the estimate $\hat{cdf}(A_{n+1}|Q_{n+1})$ is

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

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

Proof. Let us express Eq.9 with a discretization of O^j_p with prior $a^j_0 f_p \Delta p$

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

where the sum runs over all j and p by steps of Δp such that $O^j_p(A_{n+1}|Q_{n+1}) \leq x$. Since 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 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. Second, the Kolmogorov complexity of p, determining the prominent contribution of its prior, is likely chaotic and would yield very different priors than what is 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 approximated by selecting an adequate Turing machine of reference, assuming all contributions, not just the prominent ones defined by their Kolmogorov complexities, are considered, otherwise the prior will likely be confined to an exponential one, as pointed out in [2].

4.3 Operators as Beta Distributions

We have now what we need to model our rules, second order conditional probabilities, as operators.

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, when considering the rule

- *If the sun shines, then the temperature rises*

questions pertaining to what happens at night will be ignored by it.

Second, we assume that answers are Boolean, so that $A_i \in \{0, 1\}$ for any i. In reality, OpenCog rules manipulate predicates (generally fuzzy predicates but that can be let aside), and the questions they represent are: *if some instance holds property R, what are the odds that it holds property S?* We simplify this by fixing predicate S so that the problem is reduced to finding R that best predicts it. Thus we assume that if $A_i = A_{n+1}$ then O^j_p models the odds of $S(Q_i)$, and if $A_i \neq A_{n+1}$, it models the odds of $\neg S(Q_i)$. More formally, the class of operators under consideration can be represented as programs of the form

$$O^j_p(A_i|Q_i) = \text{if } R^j(Q_i) \text{ then } \begin{cases} p, & \text{if } A_i = A_{n+1} \\ 1 - p, & \text{otherwise} \end{cases} \tag{15}$$

where R^j is the conditioning predicate of the rule. This allows an operator to be modeled as a Beta distribution, with cumulative distribution function

$$cdf_{O^j}(x) = I_x(m^j + \alpha, n^j - m^j + \beta) \tag{16}$$

where m^j is the number of times $A_i = A_{n+1}$ for the subset of n^j questions such that $R^j(Q_i)$ is true. The parameters α and β are the parameters of the prior of p, itself a Beta distribution. Eq. 16 is in fact the definition of OpenCog Truth Values as described in Chapter 4 of the PLN book [5].

4.4 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 insufficiently, in our case anyway. Indeed, if a given operator cannot compute the conditional probability of some question/answer 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

[4] more by necessity, since the set of partial operators is enumerable, while the set of complete ones is not.

how to properly consider missing data [12]. Unfortunately, they do not directly apply to our case because our assumptions are different. In particular, here, data omission depends on the model. However, the general principle of modeling missing data and taking into account these models in the inference process, can be applied. Let us attempt to do that by explicitly representing the portion of the likelihood over the missing, or to use better terms, *unexplained* or *unaccounted* data of the j^{th} operator, by a dedicated term, denoted r^j. Let us also define a *completion* of O_p^j, a subprogram that explains the unaccounted data.

Definition 1. *A* completion *C of O_p^j is a program that completes O_p^j for the unaccounted data, when $R^j(Q_i)$ is false, such that the operator once completed is as follows*

$$\begin{aligned} O_{p,C}^j(A_i|Q_i) = & \textit{ if } R^j(Q_i) \textit{ then } \begin{cases} p, & \textit{if } A_i = A_{n+1} \\ 1-p, & \textit{otherwise} \end{cases} \\ & \textit{else } C(A_i|Q_i) \end{aligned} \tag{17}$$

The likelihood given the operator completed is

$$l_p^j = p^{m^j}(1-p)^{n^j - m^j} r^j \tag{18}$$

where the binomial term account for the likelihood of the explained data, and r^j accounts for the likelihood of the unexplained data, more specifically

$$r^j = \prod_{i \leq n \;\wedge\; \neg R^j(Q_i)} C^j(A_i|Q_i) \tag{19}$$

where C^j is the underlying completion of O_p^j. One may notice that r^j does not depends on p. Such assumption tremendously simplifies the analysis and is somewhat reasonable to make. We generally assume that the completion of the model is independent on its pre-existing part. By replacing the likelihood in Eq. 11 by Eq. 18 we obtain

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

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

$$f_p = f(p; \alpha, \beta) \tag{21}$$

where f is the pdf of the Beta distribution as defined 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 probably the most accepted due to being *uninformative* in some sense [10]. We do not need to commit to a particular one at that point and let the parameters α and β free, giving us

$$\hat{cdf}(A_{n+1}|Q_{n+1})(x) = \sum_j a_0^j \int_0^x \frac{p^{\alpha-1}(1-p)^{\beta-1}}{\mathrm{B}(\alpha,\beta)} p^{m^j}(1-p)^{n^j - m^j} r^j dp \tag{22}$$

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 will require normalization anyway

$$\hat{cdf}(A_{n+1}|Q_{n+1})(x) \propto \sum_j a_0^j r^j \int_0^x p^{m^j+\alpha-1}(1-p)^{n^j-m^j+\beta-1}dp \tag{23}$$

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

$$\hat{cdf}(A_{n+1}|Q_{n+1})(x) \propto \sum_j a_0^j r^j B(x; m^j+\alpha, n^j-m^j+\beta) \tag{24}$$

Using the regularized incomplete beta function we obtain

$$\hat{cdf}(A_{n+1}|Q_{n+1})(x) \propto \sum_j a_0^j r^j I_x(m^j+\alpha, n^j-m^j+\beta)B(m^j+\alpha, n^j-m^j+\beta) \tag{25}$$

As I_x is the cumulative distribution function of O^j (Eq. 16), we finally get

$$\hat{cdf}(A_{n+1}|Q_{n+1})(x) \propto \sum_j a_0^j r^j cdf_{O^j}(x)B(m^j+\alpha, n^j-m^j+\beta) \tag{26}$$

We have expressed our cumulative distribution function estimate as an averaging of the cumulative distribution functions of the operators. This gives us an estimate that predicts to what extend S holds for a new question and how much confidence we have in that prediction.

However, 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 naturally exhibit more uncertainty due to carrying less evidence.

4.5 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. Meaning, we must consider not only the complexity of the rule but also the complexity of its completion. Unfortunately calculating that complexity is intractable. To work around that we estimate it as function of the length of the unexplained data. Specifically, we suggest as prior

$$a_0^j = 2^{-K(O^j)-v_j^{(1-c)}} \tag{27}$$

where $K(O^j)$ is the Kolmogorov complexity of the j^{th} operator (the length of its corresponding rule in bits), v_j is the length of its unaccounted data, and c is a *compressability* parameter. If $c = 0$ then the unaccounted data are incompressible. If $c = 1$ 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 early stage nature and the space limitation of the paper, a value of $c = 0.5$ was actually shown to be somewhat satisfactory.

5 Conclusion

We have introduced a specialization of Solomonoff Operator Induction over operators with the particularities of being partial and modeled by Beta distributions. A second order probability estimate to predict new data, as well as capturing the uncertainty of such prediction, has been derived. While doing so we have uncovered an interesting problem, how to account for partial operators in the estimate. This problem appears to have no obvious solution, is manifestly under-addressed by the research community, and 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 provides some initial ground for experimentation and motivates further research. Even though, ultimately, it is expected that this problem might be hard enough to require some form of meta-learning [3], 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 enabling inference control meta-learning within the OpenCog framework and will be the subject of 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.: Toward a formal characterization of real-world general intelligence. Artificial General Intelligence: 3d International Conference (2010)
3. 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)
4. 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)

5. Goertzel, B., Ikle, M., Goertzel, I.F., Heljakka, A.: Probabilistic Logic Networks. Springer US (2009)
6. Hoeting, J.A., Madigan, D., Raftery, A.E., Volinsky, C.T.: Bayesian model averaging: a tutorial. Statist. Sci. **14**(4), 382–417 (1999)
7. Hutter, M.: Optimality of universal bayesian sequence prediction for general loss and alphabet. Journal of Machine Learning Research **4**, 971–1000 (2003)
8. Ikle, M., Goertzel, B.: Probabilistic quantifier logic for general intelligence: An indefinite probabilities approach. Artificial General Intelligence: First International Conference pp. 188–199 (2008)
9. J. Solomonoff, R.: Three kinds of probabilistic induction: Universal distributions and convergence theorems. Comput. J. **51**, 566–570 (2008)
10. 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)
11. Li, M., Vitanyi, P.: An introduction to Kolmogorov Complexity and its Applications. Springer-Verlag, N. Y. (1997)
12. Schafer, J.L., Graham, J.W.: Missing data: Our view of the state of the art. Psychological Methods p. 177 (2002)
13. Weisstein, E.W.: Regularized beta function. From MathWorld—A Wolfram Web Resource, http://mathworld.wolfram.com/RegularizedBetaFunction.html, last updated: Fri Apr 20 2018