Design of Stochastic Machines Dedicated to Approximate Bayesian inferences
Marvin Faix, Raphä Laurent, Pierre Bessière, Emmanuel Mazer, Jacques Droulez

To cite this version:
Marvin Faix, Raphä Laurent, Pierre Bessière, Emmanuel Mazer, Jacques Droulez. Design of Stochastic Machines Dedicated to Approximate Bayesian inferences. IEEE Transactions on Emerging Topics in Computing, Institute of Electrical and Electronics Engineers, 2016, <10.1109/TETC.2016.2609926>. <hal-01374906>

HAL Id: hal-01374906
https://hal.archives-ouvertes.fr/hal-01374906
Submitted on 2 Oct 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Design of Stochastic Machines Dedicated to Approximate Bayesian Inferences

Marvin Faix, Raphaël Laurent, Pierre Bessière, Emmanuel Mazer and Jacques Droulez

Abstract—We present an architecture and a compilation toolchain for stochastic machines dedicated to Bayesian inferences. These machines are not Von Neumann and code information with stochastic bitstreams instead of using floating point representations. They only rely on stochastic arithmetic and on Gibbs sampling to perform approximate inferences. They use banks of binary random generators which capture the prior knowledge on which the inference is built. The output of the machine is devised to continuously sample the joint probability distribution of interest. While the method is explained on a simple example, we show that our machine computes a good approximation of the solution to a problem intractable in exact inference.

Index Terms—Stochastic computing, Bayesian Programming.

1 INTRODUCTION

This project is inspired by E.T. Jaynes. His book, “Probability as Logic” [1], proposes probability as an extension of logic to reason with incomplete and uncertain information. Since the beginning of our work in the 90s [2], [3], it has always been our goal to build a machine able to perform this type of reasoning on specialized hardware. We started by designing a method to automate probabilistic reasoning called Bayesian Programming [4] and an associated programming language ProBT to specify and perform Bayesian inferences on standard computers. The formalism and the inference engine of ProBT have been used to develop many academic [5], [6], and industrial applications.

Recently the European project BAMBI gave us the opportunity to pursue our project. We started this project by designing and implementing machines performing exact inference with stochastic arithmetics on an FPGA. While these non conventional machines could already be used to synchronize a pair of Linear Feedback Shift Registers [7] or to control a robot [8], they could not handle any problem of reasonable complexity. Here we describe a second generation of machines based on sampling which allow to approximate the solution of otherwise intractable problems.

The paper is organized as follows. Related works with similar goals and approaches are described next. Section 3 presents stochastic bitstreams and explains why we made the choice to sample binary probabilistic variables. In particular, it introduces the two only building blocks required to build the stochastic machine: Conditional Distribution Elements (CDEs) and Extended C-Elements (ECEs). CDEs are used to generate programmable and addressable stochastic bitstreams. ECEs extend Muller C elements and are used for inference. In the following section we recall the Gibbs algorithm in the case of binary variables and show how it may be implemented by combining our previously defined building blocks. Then, in Section 5 we describe our compilation toolchain. It starts from a Bayesian program on discrete variables, transforms it into an equivalent program on binary variables and finally compiles it into a VHDL circuit specification. Finally we use a simulator to demonstrate the performance of the obtained circuits performing approximate inference on an intractable problem.

2 RELATED WORK

The design of unconventional or non Von Neumann machines is essentially driven by the idea of reducing the power consumption of computing devices. It is also related to the need to process larger amounts of data while being closer as ever to the limits imposed by physical laws on current processors and the end of Moore’s law. The need for more data processing with less energy comes from the common availability of extremely large data bases and from applications related to sensor fusion and interpretation.

Massively parallel and multi-core architectures are the current response to these challenges but several projects tend to investigate alternate paths. Projects like SpiNNaker and TrueNorth are still standing on the side of the multi-core approach while being bio-inspired. The global architecture they propose changes to comply with the artificial neural network inspiration, but the core computations are still made with Von Neumann processor architectures.

The PCMOS (Probabilistic-CMOS) project [9] modifies the structure of processors to build resilient and energy efficient machines. The key idea is to lower the nominal voltage at which standard CMOS circuits operate and to compensate errors with a robust design. An algebra formalizes how errors are propagated along a chain of logical

---

operations, leading to circuits computing the correct result with a specified probability. Contrary to our approach, in which we design new theory and components to compute inference, PCMOS manages errors on classic CMOS arithmetic operators to perform approximate computation with compromises between circuit size, voltage, and precision.

Closer from our own work are the approaches taken by Vigoda, Masinkha and Jonas at MIT. In all cases, probabilistic programming is seen as a way to program/specify the future machines and they all avoid the Von Neuman architecture as well as the use of FPU’s. Vigoda [10] was the first to initiate, in 2003, the design of non conventional machines dedicated to Bayesian inference. The track chosen was to represent probability distributions on binary variables with analog electrical values. He focused on a well known algorithm for exact inference: the message passing algorithm, and devised the necessary analog components to perform the message passing to obtain the circuit performing the desired inference. While the approach has many practical applications it will not handle complex inference problems.

To address this scalability issue, V. Mansinghka [11] and E. Jonas [12] are proposing a sampling approach based on stochastic calculus: Gibbs samplers or MCMC methods (Metropolis-Hastings) to perform approximate probabilistic inferences and return samples of the posterior distribution. These methods are already established in probabilistic languages such as Church [13] or Venture [14], but also exist inside inference engines supporting approximate inference such as ProBT [15]. To obtain a hardware implementation they proposed a set of transition circuits to sample quantized or continuous variables. In [16] the circuits are based on building blocks performing operations (for example normalisation) by evaluating mathematical expressions with dedicated hardware or by using precomputed tables. Our goal was to propose simpler building blocks, and the architecture we propose only uses two types of probabilistic gates: the CDE (similar to the CPT gate of Jonas) and the ECE gate which has no equivalent in the previously proposed architectures.

Our project is also strongly related to the study of nanoscale devices to perform Bayesian inference, which motivates our architecture choices. These devices are nanoscale samplers for Poisson distributions which could eventually replace the current source of entropy (PRNG for pseudo random number generator) of the presented architecture, as in the work of the Computing Fabrics Laboratory of the University of Massachusetts Amherst [17] which is presenting an unconventional hardware architecture and nanoscale technology implementation of Bayesian inference.

### 3 Stochastic Bitstreams and Operators

#### 3.1 Definition

A stream of bits may be considered as encoding the probability $p$ of a binary variable: $p = n_1/(n_1 + n_0)$ where $n_1$ is the number of 1 and $n_0$ the number of 0 in the stream. It may also be considered as coding the odds $o = p/(1-p)$ of this binary variable. The odds ratio is then given by $o = n_1/n_0$.

Conversely, given a probability $p$ (or an odds ratio $o$) we may use software or hardware stochastic processes to generate a bitstream $B(p)$ (respectively $B(o)$) that approximately encodes the probability $p$ (respectively the odds ratio $o$).

Operators on bitstreams may then be used to perform probabilistic calculus on binary distributions. A central point of this paper is to show that using only two simple operators called CDE and ECE we can solve complex probabilistic problems.

As these operators compute with an approximate representation of probabilities, we will use the root of the mean square error (RMSE) to measure the precision of an operator: if $p'$ is the observed result given by an operator and $p$ the desired result, we compute the RMSE by considering several experiments leading to several $p'$ and averaging the square error on $p$ and $1 - p$.

$$\text{RMSE} = \frac{1}{\sqrt{2}} \sqrt{p - p'}^2.$$  

#### 3.2 CDE: Conditional Distribution Element

Gupta & Kumaresan [18], [19], proposed a circuit able to generate a stochastic bitstream $B(p)$ encoding a probability $p$ stored in a fixed point format $p = \sum_{i=1}^{m} x_i 2^{-i}$ in a buffer $X$ (see red part of Figure 1).

We propose to complement this circuit by an addressable memory (see blue part of Figure 1). Given a value in $Y$ the content of the memory at this address (a given probability $p_Y$) is written in $X$. The generated stochastic bitstream $B(p_Y)$ is then conditioned by the value $Y$.

![Fig. 1. Hardware implementation of the Conditional Distribution Element.](image)

#### 3.3 ECE: Extended C-Element

As in Gaines [19], [20] we use operators to combine stochastic signals and perform arithmetic operations. For example, assuming $B(p_1)$ and $B(p_2)$ are two independent stochastic bitstreams we may use a simple AND gate to obtain $B(p_1, p_2)$. As we will see in the next section, what we really need is the product of two odds: a way to generate $B(o_1, o_2)$ from two independent bitstreams $B(o_1)$ and $B(o_2)$. For that purpose we propose to use an extension of Muller C-elements using more memory resources.

The Muller C-element truth table is given in Table 1. It uses a one bit memory to provide the previous value $Z_{prev}$.
as output when the inputs are either (0, 1) or (1, 0). When two independent and not autocorrelated bitstreams $B(o_1)$ and $B(o_2)$ are given as inputs to a $C$-element, the output $B(o_1,o_2)$ is an unbiased estimator of the product of the odds $o_1$ and $o_2$. This result can be obtained by applying the detailed balance principle to the Markov chain $Z_n \rightarrow Z_{n+1}$.

However, there is some autocorrelation in the output stream (a bit at time $t$ is not independent of a bit at time $t - 1$) due to the effect of memory. This autocorrelation induces troubles when $C$-elements are cascaded [21]. To reduce the autocorrelation of the output, we extend the memory capacity using a buffer $C$ of size $D$ bits. This toroidal buffer works as follows. When the inputs have the same value ($X = Y = v$) the content of the buffer is shifted to the left, its rightmost bit is set to $v$ and we provide $v$ as output. When the inputs have different values, we provide the rightmost value of $C$ as output and rotate the buffer content to the right. If $D=1$ we have a regular Muller $C$ element. The truth table of the $ECE$ is presented in Table 2.

Several methods using some memory to mitigate the bitstream autocorrelation issue have been compared in [22] where the authors conclude that Edge Memories provide the best performances and speed on their decoding application. The toroidal buffer of the $ECE$ serves the same purpose, but without the need of generating a random address, thus reducing the circuit area and power needs. Figure 2 presents the experimental RMSE for a cascade of two products of odds computed either with $ECE$s or with Edge Memories. As the buffer size is increased, the error decrease (until the precision threshold determined by the bitstream size $N$ is reached) is faster for $ECE$s than for Edge Memories.

### 4 Gibbs Algorithm and Product of Odds

#### 4.1 Gibbs algorithm

Let us consider a set $X$ of $|X|$ binary variables $X = \{B_1, \cdots, B_{|X|}\}$. Let us consider a subset $K = \{K_1, \cdots, K_{|K|}\}$ of $X$ made of binary variables with known values $\{k_1, \cdots, k_{|K|}\}$. Let us consider a subset $I = \{I_1, \cdots, I_{|I|}\}$ of $X$ made of binary variables of interest. Finally, let us consider a complementary subset $F$ of free variables. We have, $X = K \cup I \cup F$ and, of course, $|X| = |K| + |I| + |F|$. The purpose of the Gibbs algorithm is to sample the distribution $P(I_1, \cdots, I_{|I|}|k_1, \cdots, k_{|K|})$. It operates as follows:

**for** $B_i \in K$ **do**

Set $B_i$ to the corresponding value $k_j$

**end for**

**for** $B_i \in I \cup F$ **do**

Draw a value $b_i$ at random.

**end for**

**while** TRUE **do**

**for** $B_i \in I \cup F$ **do**

Draw $b_i$ according to $P(B_i|b_1 \ldots b_{i-1}, b_{i+1}, \ldots, b_{|X|})$

Output the binary values of the variables in $I$

**end for**

**end while**

We see that this algorithm only needs to be able to sample from a distribution $P(B_i|b_1 \ldots b_{i-1}, b_{i+1}, \ldots, b_{|X|})$. Any joint probability distribution on $|X|$ binary variables can be written as:

\[
P(B_1, B_2, \cdots, B_{|X|}) = \prod_{n=1}^{|X|} P(B_n|B_1, \cdots, B_{n-1})
\]

(1)

If we are interested in the probability distribution of one of these binary variables $B_i$ knowing the $|X| - 1$ values $b_j$ of all the others, we get:

\[
P(B_i|b_1, \ldots, b_{i-1}, b_{i+1}, \ldots, b_{|X|}) \propto \prod_{n=1}^{i-1} P(b_n|b_1, \ldots, b_{n-1}) P(B_i|b_1, \ldots, b_{i-1}) \prod_{n=i+1}^{|X|} P(b_n|b_1, \cdots, B_i, \ldots, b_{n-1})
\]

(2)

To compute the odds of $B_i$, we get:
\[
O(B | b_1, \ldots, b_{i-1}, b_{i+1}, \ldots, b_{X}) = \prod_{n=i}^{1} \left[ \frac{P(b_1 | b_2, \ldots, b_{n-1})}{P(b_1 | b_2, \ldots, b_{n-1})} \right] \cdot \prod_{n=i+1}^{X} \left[ \frac{P(b_i | b_1, \ldots, b_{i-1}, b_{i+1}, \ldots, b_{n-1})}{P(b_i | b_1, \ldots, b_{i-1}, b_{i+1}, \ldots, b_{n-1})} \right].
\]

The first term vanishes:

\[
= \frac{P(R=1 | s)}{P(R=0 | s)} \cdot \frac{P(s | R=1)}{P(s | R=0)} \cdot \frac{P(S=1 | R=1, S)}{P(S=0 | R=0)}.
\]

As some variables may be independent, it may occur that some terms in the remaining products are independent of \(R\), in which case they also get simplified. Equation 4 shows how to compute the odds of any binary variable knowing the value of all the others as a product of probability ratios. Each term has a likelihood, homogeneous to odds \(o\), which can be sampled by bitstreams generated with probability \(p = o / (1 + o)\) by CDEs described in Section 3.2, and their product can be computed by ECEs described in Section 3.3.

### 4.3 Stochastic Gibbs Circuit
#### 4.3.1 Principle
Let us take an example with three binary variables \(X = \{R, S, G\}\) to present the architecture of the stochastic Gibbs circuit. This example is inspired from the classical “sprinkler” Bayes net. \(R, S, G\) are binary variables corresponding to the predicates “it has been Raining”, “the Sprinkler was turned on” and “the Grass is wet”. The joint distribution may be decomposed as \(P(R, S, G) = P(R)P(S|R)P(G|R, S)\). We are interested in the probability that it rained knowing that the grass is either wet or dry: \(P(R|G = g)\). We have \(K = \{G\}, I = \{R\}\) and \(F = \{S\}\.

According to the Gibbs algorithm we need to sample \(R\) and \(S\), which means, according to Equation 4, to compute two odds with the following formulas:

\[
O(R | G = g, S = s) = \frac{P(R=1 | s)}{P(R=0 | s)} \cdot \frac{P(s | R=1)}{P(s | R=0)} \cdot \frac{P(g | R=1, S)}{P(g | R=0, S)},
\]

\[
O(S | G = g, R = r) = \frac{P(S=1 | r)}{P(S=0 | r)} \cdot \frac{P(g | S=1)}{P(g | S=0)}.
\]

These formulas include 13 ratios which are preliminary knowledge specified in the model. We use CDEs to store them and to produce the associated bitstreams. We need one CDE for each of the 3 terms appearing in the decomposition. One for \(P(R)\) to produce when needed a bitstream \(B(o_R)\) with \(o_R = \frac{P(R=1)}{P(R=0)}\). One for \(P(S|R)\) with a 4 slot memory to produce respectively \(B(o_S^1)\) with \(o_S^1 = \frac{P(S=1 | R=1)}{P(S=0 | R=1)}\), \(B(o_S^2)\) with \(o_S^2 = \frac{P(S=1 | R=1)}{P(S=0 | R=0)}\), \(B(o_S^3)\) with \(o_S^3 = \frac{P(S=0 | R=1)}{P(S=0 | R=0)}\) and \(B(o_S^4)\) with \(o_S^4 = \frac{P(S=1 | R=1)}{P(S=1 | R=0)}\). Finally, one more for \(P(G|R, S)\) with a 8 slot memory.

We then need 2 ECEs to compute the product of the 3 ratios required to sample \(o(R | G = g)\) or the 2 required for \(o(S | G = g)\). This leads to the circuit sketched in Figure 3 dedicated to computing the inference \(P(R | G = g)\). The input has to be set to the desired value \(g\) and the output is a binary bitstream \(B(o)\) encoding \(O(R | G = g)\).

### 5 EXPERIMENTATION AND RESULTS
#### 5.1 The compilation toolchain
A compilation toolchain was developed to evaluate the proposed architecture. The input is any Bayesian program using discrete variables written in ProBT (see Figure 4 for an example). The output is a VHDL program specifying the circuit dedicated to the corresponding inference.

#### 5.2 Preprocessing
Any probabilistic program using discrete variables is first transformed into an equivalent specification using only binary variables. For example, a distribution \(P(V)\) on a discrete variable with \(2^n\) values will be transformed into one distribution and \(n - 1\) conditional distributions on binary variables \(B_i^V : i \in 1 \ldots n\).
\[ P(V) = P(B_1^V, B_2^V, B_3^V, \ldots, B_{n-1}^V) = P(B_1^V)P(B_2^V | B_1^V)P(B_3^V | B_2^V, B_1^V) \ldots P(B_{n-1}^V | B_{n-2}^V, \ldots, B_1^V). \]

While the “Binary” representation looks less compact, it requires at most the same number of parameters as for representing \( P(V) \). Each probability distribution, as for instance \( P(B_2^V | B_1^V) \), is automatically computed using the inference engine of ProBT.

### 5.3 Compiler

The compiler starts from the previously generated “binary” Bayesian program and generates a VHDL file describing the final circuit. It specifies a circuit implementation of a Bayesian program and generates a VHDL file describing the associated logic circuit. The compiler starts from the previously generated “binary” representation (see the Appendix) and choses the parametric forms so that we could evaluate the quality of the analytic approximation to evaluate the results since the exact inference solution is out of reach. The precision depends on:

- The size \( D \) of the buffers used in ECEs to prevent the output bitstream autocorrelation.
- The burn-in duration \( BI \) allowing the ECEs to stabilize the output bitstream encoding the product \( B(o_1o_2) \).
- The number \( N \) of samples chosen to approximate the output distribution.

Figure 5 shows the histogram resulting from the approximate inference of \( P(V \mid b_1 = 1, \ldots, b_{20} = 1) \) with the following parameters: \( D = 16 \), \( BI = 200 \), \( N = 100000 \). The size \( D \) of the ECE buffers depends on the depth of ECE cascades. In the example Figure 2, the precision was evaluated with a two stage product of odds. Here we use ECEs with a buffer size \( D=16 \), which provides a good compromise between size and precision for this example cascading 3 ECEs. (Indeed, we compute the product of 8 odd ratios, by arranging the ECEs as a binary tree of depth 3.)

Table 3 shows the impact of the burn-in \( BI \) and of the number \( N \) of samples on the Kullback-Leibler divergence between the analytic approximation and the results obtained with our stochastic architecture. We observe that the precision increases with \( N \), but slowly. This is unsurprising since as mentioned before the precision of Bernoulli sequences go as \( \sqrt{N} \). Increasing the length of the burning sequence after which the samples computed by the tree of ECEs is considered valid has a cost in terms of computation time, but it allows to significantly increase the precision.
of the results until they are as good as for floating point computations. This suggests a way to use the same circuits with different tradeoffs between latency (and power needs) and the desired precision.

<table>
<thead>
<tr>
<th>Number of samples</th>
<th>burn in = 10</th>
<th>burn in = 20</th>
<th>burn in = 50</th>
<th>burn in = 100</th>
<th>burn in = 200</th>
<th>burn in = 400</th>
<th>floating point</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1.5293</td>
<td>1.3145</td>
<td>0.2497</td>
<td>0.1414</td>
<td>0.1335</td>
<td>0.1529</td>
<td>0.1458</td>
</tr>
<tr>
<td></td>
<td>1.2222</td>
<td>1.1482</td>
<td>0.1000</td>
<td>0.0178</td>
<td>0.0155</td>
<td>0.0151</td>
<td>0.0184</td>
</tr>
<tr>
<td></td>
<td>1.15294</td>
<td>1.13688</td>
<td>0.07979</td>
<td>0.00218</td>
<td>0.00182</td>
<td>0.00188</td>
<td>0.00181</td>
</tr>
</tbody>
</table>

TABLE 3
Kullback divergence from the Closed-Form Solution (in bits).

6 CONCLUSION
We have presented a compilation toolchain to design automatically stochastic machines that solve approximate Bayesian inference problems using the Gibbs sampling algorithm. This compilation toolchain is generic: it can translate any inference problem expressed as a Bayesian program into a circuit comprised only of two types of stochastic components: the CDE and the ECE. We implemented a quantitative test to evaluate our architecture on intractable problems. The results are promising: the output distribution of our stochastic machine approximates the analytic solution as well as a floating point implementation of Gibbs sampling.

In the near future we will extend and test our approach on Bayesian filters. In a second step, we will design a generic and programmable machine, which will accept any new Bayesian program without the need for a new circuit synthesis, opening the way to actually programmable Bayesian computing devices.

ACKNOWLEDGMENTS
This work was made possible thanks to two grants from CNRS: Nano-Bayes and Defi-Bayes, and thanks to the EU collaborative FET Project BAMBI FP7-ICT-2013-C, project number 618024.

REFERENCES
Marvin Faix (born in 1990) received the M.Eng degree in Integrated Electronic Systems in 2013 from the Ecole Nationale Supérieure de Physique, Electronique and Matériaux (PHELMA), Grenoble, France. He started his PhD in 2014 on the research topic "Probabilistic computing: from theory to hardware implementation", granted by the European FET Project BAMBI – Bottom-up Approaches to Machines dedicated to Bayesian Inference – (FP7-ICT-2013-618024). He works at the Laboratoire Informatique de Grenoble (LIG) and in collaboration with the laboratory TIMA (Techniques de l'Informatique et de la Microélectronique pour l'Architecture des systèmes intégrés). His work is oriented at probability theory to design new ways to compute Bayesian inference in hardware.

Raphaël Laurent (born in 1988) received a computer science engineering master degree from ENSIMAG and an Artificial Intelligence master degree from the Joseph Fourier University (Grenoble, France). During his PhD at Grenoble-Alpes University he acquired valuable experience in probabilistic information processing systems as he implemented Bayesian models and algorithms to study cognitive aspects of speech perception, production and learning. He is now working in the ProbaYes R&D group which develops efficient solutions in the fields of machine learning, probabilistic models, optimization and data science for a wide range of industrial applications. His current research interests within the BAMBI project (Bottom-up Approaches to Machines dedicated to Bayesian Inference) are the design and hardware implementation of unconventional computing architectures along with the development of software tools allowing to automatically map probabilistic programs into efficient circuit implementations based on new computation paradigms.

Pierre Bessière was born in 1958. He received the engineering degree and the Ph.D. degree in computer science from the Institut National Polytechnique de Grenoble (INPG), France, in 1981 and 1983, respectively. He did a Post-Doctorate at SRI International (Stanford Research Institute) working on a project for National Aeronautics and Space Administration (NASA). He then worked for five years in an industrial company as the leader of different artificial intelligence projects. He came back to research in 1989. He has been a senior researcher at Centre National de la Recherche Scientifique (CNRS) since 1992. His main research concerns have been evolutionary algorithms and probabilistic reasoning for perception, inference and action. He leads the Bayesian Programming research group (Bayesian-Programming.org) on these two subjects. Twenty PhD diplomas and numerous international publications are fruits of the activity of this group during the last 15 years. He also leads the BIBA (Bayesian Inspired Brain and Artefact) and BAMBI (Bottom-up Approaches to Machines dedicated to Bayesian Inference) European projects and was a Partner in BACS (Bayesian Approach to Cognitive Systems). He is a co-founder and scientific adviser of the ProbaYes Company, which develops and sells Bayesian solutions for the industry.