Reactive probabilistic programming
Guillaume Baudart, Louis Mandel, Eric Atkinson, Benjamin Sherman, Marc Pouzet, Michael Carbin

To cite this version:
Guillaume Baudart, Louis Mandel, Eric Atkinson, Benjamin Sherman, Marc Pouzet, et al.. Reactive probabilistic programming. PLDI 2020 - 41th ACM SIGPLAN International Conference in Programming Language Design and Implementation, Jun 2020, London / Virtual, United Kingdom. 10.1145/3385412.3386009. hal-03051954

HAL Id: hal-03051954
https://inria.hal.science/hal-03051954
Submitted on 10 Dec 2020

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.
Reactive Probabilistic Programming

Guillaume Baudart
MIT-IBM Watson AI Lab,
IBM Research
USA

Louis Mandel
MIT-IBM Watson AI Lab,
IBM Research
USA

Eric Atkinson
MIT
USA

Benjamin Sherman
MIT
USA

Marc Pouzet
École Normale Supérieure,
PSL Research University
France

Michael Carbin
MIT
USA

Abstract
Synchronous modeling is at the heart of programming languages like Lustre, Esterel, or SCADE used routinely for implementing safety critical control software, e.g., fly-by-wire and engine control in planes. However, to date these languages have had limited modern support for modeling uncertainty – probabilistic aspects of the software’s environment or behavior – even though modeling uncertainty is a primary activity when designing a control system.

In this paper we present ProbZelus the first synchronous probabilistic programming language. ProbZelus conservatively provides the facilities of a synchronous language to write control software, with probabilistic constructs to model uncertainties and perform inference-in-the-loop.

We present the design and implementation of the language. We propose a measure-theoretic semantics of probabilistic stream functions and a simple type discipline to separate deterministic and probabilistic expressions. We demonstrate a semantics-preserving compilation into a first-order functional language that lends itself to a simple presentation of inference algorithms for streaming models. We also redesign the delayed sampling inference algorithm to provide efficient streaming inference. Together with an evaluation on several reactive applications, our results demonstrate that ProbZelus enables the design of reactive probabilistic applications and efficient, bounded memory inference.

CCS Concepts: • Theory of computation → Streaming models; • Software and its engineering → Data flow languages.

Keywords: Probabilistic programming, Reactive programming, Streaming inference, Semantics, Compilation

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

PLDI ’20, June 15–20, 2020, London, UK
© 2020 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-7613-6/20/06.
https://doi.org/10.1145/3385412.3386009

1 Introduction
Synchronous languages [2] were introduced thirty years ago for designing and implementing real-time control software. They are founded on the synchronous abstraction [4] where a system is modeled ideally, as if communications and computations were instantaneous and paced on a global clock. This abstraction is simple but powerful: input, output and local signals are streams that advance synchronously and a system is a stream function. It is at the heart of the data-flow languages Lustre [20] and SCADE [13]; it is also the underlying model behind the discrete-time subset of Simulink.

The data-flow programming style is very well adapted to the direct expression of the classic control blocks of control engineering (e.g., relays, filters, PID controllers, control logic), and a discrete time model of the environment, with the feedback between the two. For example, consider a backward Euler integration method defined by the following stream equations and its corresponding implementation in Zelus [7], a language reminiscent of Lustre:

\[ x_0 = x_0 \quad x_n = x_{n-1} + x'_n \times h \quad \forall n \in \mathbb{N}, n > 0 \]

\[
\text{let node integr (xo, x') = x where}
\text{rec x = xo -> (pre x + x' * h)}
\]

The node integr is a function from input streams xo and x' to output stream x. The initialization operator -> returns its left-hand side value at the first time step and its right-hand side expression on every time step thereafter. The unit-delay operator pre returns the value of its expression at the previous time step. The following table presents a sample timeline showing the sequences of values taken by the streams defined in the program (where h is set to 0.1).
The node `integ` can be used to define other stream functions, e.g., a PID controller, which can be called in control structures like hierarchical automata, e.g., to express a system that switches between automatic and manual control. Compared to a general purpose functional language (or an embedded DSL), the expressiveness of a synchronous language is purposely constrained to modularly ensure safety properties that are critical for the targeted applications: determinacy, deadlock freedom (reactivity), the generation of statically scheduled code that runs in bounded time and space.

However, to date, these languages have had limited support for modeling uncertainty (e.g., a noisy sensor or channel, a variable delay), to simulate the interaction of a software controller and a partially unknown environment, or to infer parameters from noisy observations. Moreover, uncertainty is a first-order design concern for a controller that operates under the assumption of a probabilistic model of their environment (e.g., object tracking). Using a probabilistic environment model and data gathered from observing the environment, a controller can infer a distribution of likely environments given the observations. Existing approaches consist in hand-coding stochastic controllers that have a known solution (e.g., Kalman filters) which can be tedious and error-prone, or to simply perform off-line statistical testing on the generated code of a controller. Alternatively, in recent years, probabilistic programming has developed as an approach to endow general purpose languages with the ability to automate inference.

**Probabilistic Programming.** Probabilistic programming languages are used to describe probabilistic models and automatically infer distributions of latent (i.e., unobserved) parameters from observations.

A popular approach [6, 18, 28, 36–38] consists in extending a general-purpose programming language with three constructs: (1) `x = sample(d)` introduces a random variable `x` of distribution `d`, (2) `observe(d, y)` conditions on the fact that a sample from distribution `d` produced the observation `y`, and (3) `infer m obs` computes the distribution of the output values of a program or model `m` w.r.t. the observation of the input data `obs`.

Probabilistic programming languages offer a variety of automatic inference techniques ranging from exact inference [17] to approximate inference [29] and include hybrid approaches that combine exact and approximate techniques when part of the program is analytically tractable [27]. However, a standing challenge for these programming languages is that none of them meet the design goals of synchronous reactive languages by being immediately amenable to techniques to ensure that for example a program with an indefinite execution time runs in bounded memory.

**Inference in the Loop.** In this paper we extend Zelus to provide a synchronous probabilistic programming language, ProbZelus. ProbZelus enables one to combine deterministic reactive data-flow programs, such as `integ` (above), with probabilistic programming constructs to produce reactive probabilistic programs.

Compared to other probabilistic languages (e.g. WebPPL, Church, Stan) where inference is executed on terminating pure functions, our probabilistic models are stateful stream processors. Inference on probabilistic models runs in parallel with deterministic processes that interact with the environment. The distributions computed by `infer` at each step can thus be used by deterministic processes to compute new inputs for the next inference step. We term this capability **inference-in-the-loop**.

**Streaming Inference.** ProbZelus provides multiple inference algorithms, most notably the delayed sampling inference algorithm [27]. This hybrid strategy combines the approximate inference technique of particle filtering [19] with exact inference when it is possible to symbolically determine the exact distribution for some or all of the latent variables of the program [16].

However, the memory consumption of delayed sampling strictly increases with the number of random variables which is not practical for reactive applications that operate on infinite streams. We propose a novel streaming implementation of delayed sampling that can operate over infinite streams in constant memory for a large class of models. ProbZelus therefore provides an expressive language for reactive probabilistic programming with appropriate memory consumption properties.

**Contributions.** We present the following contributions:

**Design, Semantics, Compilation.** We present ProbZelus, the first synchronous probabilistic programming language, combining language constructs for streams (reactivity) with those for probabilistic programming thus enabling inference-in-the-loop. We give a measure-based co-iterative semantics for ProbZelus that forms the basis of a compiler and demonstrate a semantics-preserving compilation strategy to a first-order functional language µF.

**Streaming Inference.** We define the semantics of multiple inference algorithms on µF including particle filtering and delayed sampling. We then present a novel streaming delayed sampling implementation which enables partial exact inference over infinite streams in bounded memory for a large class of models.

---

We consider a robot equipped with an accelerometer and a GPS. We assume that the motion of the robot can be described as: $x_{t+1} = Ax_t + Bu_t$ where $x_t$ denotes the state of the robot (position, velocity, and acceleration) at a given time step $t$, and $u_t$ denotes the command sent to the robot. $A$ and $B$ are constant matrices. In addition, the robot receives at each step noisy observations from the accelerometer $a_t$, and sporadically an estimation of the position from a GPS $p_t$.

Figure 1 presents a controller, robot, that given an initial state $x_0$, an initial command $u_0$, and inputs from the accelerometer $a$ and the GPS $p$, computes a stream $u$ of commands that drives the robot to a given target. The body of robot is the parallel composition of (i) the inference of a probabilistic process $\text{kalman}$ that estimates $x_{\text{dist}}$ the stream of distributions over the robot’s state, and (ii) a deterministic process that computes the stream $u$ of commands. It is written as two mutually recursive equations that define $x_{\text{dist}}$ from $u$ and $u$ from the previous value of $x_{\text{dist}}$. The command $u$ is set to the initial command $u_0$ at the first time step, and is then computed by a Linear-Quadratic Regulator (LQR) [34] — a stable and optimal controller for such dynamic systems — given the estimation of the state at the previous step. Because LQR controllers depend only on mean posterior state, the example in Figure 1 uses the mean function to compute the mean of $x_{\text{dist}}$ before invoking the LQR controller.

**Evaluation.** We evaluate the performance of ProbZelus on a set of benchmarks that illustrate multiple aspects of the language. We demonstrate that streaming delayed sampling drastically reduces the number of particles required to achieve better accuracy compared to a particle filter.

The result is ProbZelus, a synchronous probabilistic language that enables us to write, in the very same source, a deterministic model of the control software and a probabilistic model with complex interactions between the two. On one hand, a deterministic model of a controller can rely on predictions computed by a probabilistic model. On the other hand, a probabilistic model can be programmed in an expressive reactive language. ProbZelus is open source (https://github.com/IBM/probzelus). An extended version with appendices of the paper is also available [1].

### 2 Example

In this section, we demonstrate how ProbZelus provides probabilistic modeling, inference-in-the-loop, and bounded-memory inference for a robot navigation system. The results of the inference are continuously used by a controller to correct the robot trajectory.

#### 2.1 Inference in the Loop.

We consider a robot equipped with an accelerometer and a GPS. We assume that the motion of the robot can be described as: $x_{t+1} = Ax_t + Bu_t$ where $x_t$ denotes the state of the robot (position, velocity, and acceleration) at a given time step $t$, and $u_t$ denotes the command sent to the robot. $A$ and $B$ are constant matrices. In addition, the robot receives at each step noisy observations from the accelerometer $a_t$, and sporadically an estimation of the position from a GPS $p_t$.

Figure 2 presents a Kalman filter for the robot example. Variables are either latent (white, e.g., state $x$) or observed (gray, e.g., acceleration $a$). The position $p$ is only sporadically observed.

![Kalman Filter](https://example.com/kalman.png)

**Figure 2.** Kalman filter for the robot example. Variables are either latent (white, e.g., state $x$) or observed (gray, e.g., acceleration $a$). The position $p$ is only sporadically observed.

**Sampling.** Inside the $\text{kalman}$ node, the $\text{sample}$ operator samples a value from a probability distribution. In this case, the expression samples the current state $x$ from a multivariate Gaussian with mean obtained by applying the motion model to the previous state and the command. This code models the trajectory of a robot where at each time step, the state is Gaussian-distributed around an estimation computed from the motion model.

**Observations.** The expression $\text{observe}$ conditions the execution on observed data. Its first parameter denotes a distribution that models the observation and its second parameter denotes the observed value itself. In this case, at each step, the first $\text{observe}$ statement models a Gaussian-distributed observation of the current acceleration $a_{\text{dist}}$ given by $a_{\text{dist}}$. The input $a$ is a signal that is only emitted when the

```haskell
let proba kalman (xo, u, acc, gps) = x where
  rec mu = xo -> (a => pre x) +\theta (b +\theta u)
  and x = sample (mv_gaussian (mu, noise))
  and () = observe (gaussian (vec_get (x, 2), 1.0), acc)
  and () = present gps(pos) ->
    observe (gaussian (vec_get (x, 0), 0.01), pos)
  else ()

let node robot (xo, uo, acc, gps) = u where
  rec x_dist = infer 1000 kalman (xo, u, acc, gps)
  and u = uo -> lqr a b (mean (pre x_dist))
```

![Robot controller with inference-in-the-loop](https://example.com/robot.png)

**Figure 1.** Robot controller with inference-in-the-loop. +@ and *@ are matrix operations, vec_get x i is the i-th projection.

The result is ProbZelus, a synchronous probabilistic language that enables us to write, in the very same source, a deterministic model of the control software and a probabilistic model with complex interactions between the two. On one hand, a deterministic model of a controller can rely on predictions computed by a probabilistic model. On the other hand, a probabilistic model can be programmed in an expressive reactive language. ProbZelus is open source (https://github.com/IBM/probzelus). An extended version with appendices of the paper is also available [1].
A classic operational interpretation of a probabilistic model is an importance sampler that generates random samples from the model together with an importance weight measuring the quality of the sample. In this model, each execution of a sample operator samples a value from the operator’s corresponding distribution. Each execution of an observe evaluates the likelihood of the provided observation and multiplies the current importance weight by this value. Then, each execution step of infer yields a distribution represented as a set of pairs (output, weight) or particles. The particles can be re-sampled at each step to build a particle filter [15].

The integer parameter to infer determines how many particles to use: the more particles the user specifies, the more accurate the estimate of the distribution becomes. The PF points in Figure 3 present this improvement in accuracy as a function of increasing the number of particles for the robot example. However, as the latency results presents, the more particles the user specifies, the more computation is required for each step because each particle requires a full, independent execution of each time step of the model.

Streaming Delayed Sampling. Delayed sampling [27] can reduce the number of particles required to achieve a given desired quality of inference. Specifically, delayed sampling exploits the opportunity to symbolically reason about the relationships between random variables to compute closed-form distributions whenever possible. To capture relationships between random variables, delayed sampling maintains a graph: a Bayesian network that can be used to compute closed-form distributions involving subsets of random variables. For instance, this inference scheme is able to compute the exact posterior distribution for our robot example. The SDS dots in Figure 3 show that the accuracy is independent of the number of particles since each particle computes the exact solution.

GPS computes a new position. When a value pos is emitted on gps, the present construct executes its left branch, further conditioning the model by adding a Gaussian-distributed observation of the current position vec_get x 0 given by pos.

2.2 Streaming Inference

A classic operational interpretation of a probabilistic model is the particle-to-use: the more particles the user specifies, the more computation is required for each step because each particle requires a full, independent execution of each time step of the model.

Streaming Delayed Sampling. Delayed sampling [27] can reduce the number of particles to use: the more particles the user specifies, the more computation is required for each step because each particle requires a full, independent execution of each time step of the model.

Streaming Delayed Sampling. Delayed sampling [27] can reduce the number of particles required to achieve a given desired quality of inference. Specifically, delayed sampling exploits the opportunity to symbolically reason about the relationships between random variables to compute closed-form distributions whenever possible. To capture relationships between random variables, delayed sampling maintains a graph: a Bayesian network that can be used to compute closed-form distributions involving subsets of random variables. For instance, this inference scheme is able to compute the exact posterior distribution for our robot example. The SDS dots in Figure 3 show that the accuracy is independent of the number of particles since each particle computes the exact solution.

Figure 4 illustrates the evolution of the delayed sampling graph as it proceeds through the first four time steps of the robot example (for simplicity we assume that there is no GPS activation in these four steps). A notable challenge with the traditional delayed sampling algorithm is that the graph grows linearly in the number of samples. This property is not tractable in our reactive context because would like to deploy our programs under the model that they run indefinitely, thus requiring that they execute with bounded resources. To address this problem, we propose a novel streaming delayed sampling (SDS) implementation of the delayed sampling algorithm. Specifically, in Figure 4 the node denoting the marginal posterior for x at step 1 can be eliminated from the graph at step 3 because the distributions for pre x and x have fully incorporated its effect on their values and, moreover, the program no longer maintains a reference to the node.

While the standard delayed sampling algorithm will keep this node alive through the edge pointers it maintains, SDS builds a pointer-minimal graph representation with a minimal number of edges that 1) ensure that the graph has sufficient connectivity to support operations in the traditional delayed sampling algorithm and 2) only maintain the reachability of nodes that can effect the distribution of future nodes in the graph. The result is that the memory consumption of SDS is constant across the number of steps while the memory consumption of the original delayed sampling implementation DS increases linearly in the number of steps (Figure 5).

3 Language: Syntax, Typing, Semantics

ProbZelus is a reactive probabilistic language with inference-in-the-loop which enables interaction between probabilistic models and deterministic processes. This capability introduces two design requirements. First, a probabilistic model must be able to receive inputs from an evolving environment. Second, instead of awaiting the final result of the inference, deterministic processes running in parallel need access to intermediate results. The resulting inference-in-the-loop enables feedback loops between inferred distributions from probabilistic models and deterministic processes, which our
We focus on the following kernel of ProbZelus. The missing words in the heap per steps for the robot example.

A set of equations $E$ is either an equation $x = e$ that define $x$ with the stream $e$, the initialization of a variable with a constant $\text{init } x = c$, or parallel composition of sets of equations.

Operators (op) include boolean and arithmetic operators. In addition, ProbZelus offers a library of dedicated operators to analyze distributions, such as mean and variance, that can be used in any context (probabilistic or deterministic), e.g., on the result of the inference.

The control structure $\text{present } e \rightarrow e_1 \text{ else } e_2$ is an activation condition that executes the expression $e_1$ only when the value of $e$ is true and executes $e_2$ otherwise. It differs from $\text{if } e \text{ then } e_1 \text{ else } e_2$, where both $e_1$ and $e_2$ are computed at each step (making their internal states evolve) and the returned value is chosen based on the value of $e$.

In the following example, $o1$ and $o2$ are different streams:

```
let node \text{cpt } () = 0 \text{ where rec } o = 0 \rightarrow \text{pre } o + 1

let node \text{present } vs \_ \text{if } (b) = (o1, o2) \text{ where}
rec o1 = \text{present } (b) \rightarrow \text{cpt } () \text{ else } 0
and o2 = \text{if } b \text{ then } \text{cpt } () \text{ else } 0

\text{b} \begin{array}{cccccccccc}
\text{true} & \text{true} & \text{false} & \text{false} & \text{false} & \text{true} & \text{false} & \text{true} & \ldots
\text{o1} & 0 & 1 & 0 & 2 & 0 & 0 & 3 & \ldots
\text{o2} & 0 & 1 & 0 & 3 & 0 & 0 & 6 & \ldots
\end{array}
```

The return $e_1$ every $e$ construct re-initializes the values of the $\text{init}$ equations and the corresponding $\text{last}$ expressions in $e_1$ each time $e$ is true.

The language is extended with the classic probabilistic expressions: $\text{sample}$ to draw from a distribution, $\text{observe}$ to condition on observations, and $\text{infer}$ to compute the distribution described by a model.

**Scheduling.** In the expression $e \text{ where rec } E, E$ is a set of mutually recursive equations. In practice, a scheduler orders the equations according to their dependencies. Initializations $\text{init } x_j = c_j$ are grouped at the beginning, and an equation $x_j = e_j$ must be scheduled after the equation $x_j = e_j$ if the expression $e_j$ uses $x_j$ outside a last construct. A program satisfying this partial order is said to be scheduled. The compiler can also introduce additional equations to relax the scheduling constraints and rejects programs that cannot

\[^2\text{The if construct can thus be considered as an external operator.}\]
be statically scheduled [5]. After scheduling, the expression $e$ where $rec$ $f$ has the following form.

$e$ where $rec$ init $x_1 = c_1$ ... and init $x_k = c_k$ and $y_1 = e_1$ ... and $y_n = e_n$

For simplicity, we also assume that every initialized variable is defined in a subsequent equation, i.e., $\{x_i\}_{i=1}^k \cap \{y_j\}_{j=1}^n = \emptyset$. If it is not the case, in this kernel we can always add additional equations of the form $x_j = \text{last } x_i$.

**Kernel.** All ProbZelus programs can be encoded in this kernel language. For instance, the program $\text{integ}$ of Section 1 can be rewritten in the kernel as follows:

let $\text{node}$ integ $(x_0, x') = x$ where
rec init first = true and init $x = 0.$
and first = false
and $x = \text{if last first then } x_0 \text{ else last } x + (x' \times h)$

A stream $\text{first}$ is defined such that $\text{last first}$ is only true at the first step. The $\Rightarrow$ operator is then compiled into an $\text{if}$ statement. The $\text{pre}$ operator is compiled into a $\text{last}$ operator. The initialization value is arbitrary, the compiler’s initialization analysis guarantees that this value is never used [14]. Similarly, other constructs like hierarchial automata can be re-written using $\text{present}$ and $\text{reset}$ [12].

### 3.2 Typing: Deterministic vs. Probabilistic

Deterministic and probabilistic expressions have distinct interpretations. A dedicated type system distinguishes between the two kinds of expressions, assigning one of two *kinds* to each expression: D for deterministic, or P for probabilistic. The typing judgment $G \vdash^k e : T$ states that in the environment $G$ which maps variable names to their type, the expression $e$ has kind $k$ and type $T$. Function types $T \rightarrow^k T'$ are extended with the kind $k$ of the body and we introduce a new datatype $T \text{ dist}$ for the probability distribution over values of type $T$.

The expressions $\text{sample}$, and $\text{observe}$ are probabilistic but their arguments must be deterministic. Any deterministic expression can be lifted to a probabilistic expression using a sub-typing rule. The transition from probabilistic to deterministic is realized via $\text{infer}$ (the complete set of typing rules is presented in Figure 12 of Appendix A.1).

\[
\begin{align*}
G \vdash^D e : T \text{ dist} & \quad G \vdash^D e_1 : T \text{ dist}^* \quad G \vdash^D e_2 : T \\
G \vdash^P \text{sample}(e) : T & \quad G \vdash^P \text{observe}(e_1, e_2) : \text{unit} \\
G \vdash^D e : T & \quad G \vdash^P e : T \\
G \vdash^P e : T & \quad G \vdash^D \text{infer}(e) : T \text{ dist}^* \\
G \vdash^D e : T \text{ dist}^* & \quad G \vdash^D e : T \text{ dist}
\end{align*}
\]

The type $T \text{ dist}^*$ is a subtype of $T \text{ dist}$ that represents distributions known to have a density, i.e., discrete distributions (w.r.t. the counting measure) and a subset of continuous distributions (w.r.t. the Lebesgue measure). In ProbZelus, to simplify the semantics of the language, the observe statement requires a value of type $T \text{ dist}^*$. In Appendix A, we extend the language with the $\text{factor}$ statement for arbitrary conditioning.

### 3.3 Co-Iterative Semantics

We now give the semantics of ProbZelus in a co-iteration framework [11]. In this framework, a *deterministic stream* of type $T$ is defined by an initial state of type $S$ and a transition function of type $S \rightarrow T \times S$.

$$\text{CoStream}(T, S) = S \times (S \rightarrow T \times S)$$

Repeatedly executing the transition function from the initial state yields a stream of values of type $T$.

The semantics of a *deterministic expression* ($\text{kindOf}(e) = D$) is defined using two auxiliary functions. If $\gamma$ is an environment mapping variable names to values, $\ll e \gg_{\gamma}$ denotes the initial state, and $\ll e \gg_{\gamma}$ denotes the transition function:

$$\ll e \gg_{\gamma} : \text{CoStream}(T, S) = \ll e \gg_{\gamma}, \ll e \gg_{\gamma}$$

The deterministic semantics of ProbZelus presented in Figure 6 is an extension of [11] with the control structures $\text{present}$ and $\text{reset}$ (see also Figure 13 of Appendix A.2).
The transition function of a variable always returns the corresponding value stored in the environment \( y \).

The \texttt{present} \( e \rightarrow e_1 \text{ else } e_2 \) construct introduced in Section 2 returns the value of \( e_1 \) when \( e \) is true and the value of \( e_2 \) otherwise. The state \((s_1, s_2, s_3)\) stores the state of the three sub-expressions. The transition function lazily executes the expression \( e_1 \) or \( e_2 \) depending on the value of \( e \) and returns the updated state.

The state of a set of scheduled locally recursive definitions \( e \) where \texttt{rec} \( E \) comprises three parts: the value of the local variables at the previous step which can be accessed via the \texttt{last} operator, the state of the defining expressions, and the state of expression \( e \). The initialization stores the initial values introduced by \texttt{init} and the initial states of all sub-expressions. The transition function incrementally builds the local environment defined by \( E \). First the environment is populated with a set of fresh variables \( x_1, \text{last} \) initialized with the values stored in the state that can then be accessed via the \texttt{last} operator. Then the environment is extended with the definition of all the variables \( y_i \) by executing all the defining expressions (where \( \{x_1\}_{1..k} \cap \{y_j\}_{1..n} = \{x_1\}_{1..k} \)). Finally, the expression \( e \) is executed in the final environment. The updated state contains the value of the initialized variables defined in \( E \) that will be used to start the next step, and the updated state of the sub-expressions.

**Probabilistic extension.** The semantics of a probabilistic expression (kindOf(e) = \( P \)) follows the same scheme, but the transition function returns a \textit{measure} over the set of possible pairs (result, state):

\[
\text{CoPStream}(T, S) = S \times (S \rightarrow (\Sigma_{T \times S} \rightarrow [0, \infty]))
\]

A measure \( \mu \) associates a positive number to each measurable set \( U \in \Sigma_{T \times S} \) where \( \Sigma_{T \times S} \) denotes the \( \Sigma \)-algebra of \( T \times S \), i.e., the set of measurable sets over pairs (result, state). We use the following notation for the semantics of a probabilistic expression \( e \):

\[
\mathcal{e} ： \text{CoPStream}(T, S) = \{ e \}^\mu_T, \{ e \}^\mu_S
\]

The semantics of probabilistic expressions is presented in Figure 7 (the complete semantics is in Figure 14 of Appendix A.2). This measure-based semantics is adapted from [35] to explicitly handle the state of the transition functions.

First, any deterministic expression can be lifted as a probabilistic expression. The transition function returns the Dirac delta measure \( \delta_x(U) = 1 \) if \( x \in U \), 0 otherwise on the pair returned by the deterministic transition function applied on the current state: \( \{ e \}^\mu_T(s) : T \times S \).

The probabilistic operator \texttt{sample}(e) evaluates \( e \) which returns a distribution \( \mu : T \rightarrow \text{dist} \) and a new state \( s' \rightarrow S \), and returns a measure over the pair (result, state) where the state is fixed to the value \( s' \). \texttt{observe}(e_1, e_2) evaluates \( e_1 \) and \( e_2 \) into a distribution with density \( \mu : T \rightarrow \text{dist}^* \) and a value \( v : T \), and weights execution paths using the likelihood of \( v \) w.r.t. \( \mu \) (\( \mu_{pdf} \) denotes the density function of the distribution \( \mu \)).
with probabilistic operators. We call an anonymous function, or one of the probabilistic operators `sample`, `observe`, or `infer`. The `infer` operator is tailored for ProbZelus and always takes two arguments: a transition function of the form `fun x -> e`, and a distribution of states. This operator computes the distribution of results and the distribution of possible next steps. A type system similar to the one of ProbZelus is used to distinguish deterministic from probabilistic expressions (see Appendix B.1).

### 4.2 Compilation to \( \mu F \)

The compilation function \( C \) generates a function that closely follows the transition function defined by the co-iterative semantics presented in Section 3.3. Each expression is compiled into a function of type \( S \rightarrow T \times S \) which gives a state returns a value and an updated state (see Appendix C for the complete definition). The compilation of `present` is thus:

\[
C(\text{present } e -> e_1 \text{ else } e_2) = \text{fun } (s,s1,s2) -> \\
\text{let } v, z_s' = C(e)(s) \text{ in } \\
\text{if } v \text{ then let } v_1, s_1' = C(e_1)(s1) \text{ in } (v_1,(s',s1',s2)) \\
\text{else let } v_2, s_2' = C(e_2)(s2) \text{ in } (v_2,(s',s1,s2'))
\]

The probabilistic operators `sample`, and `observe` are treated as external operators. The compilation generates code that simply calls the \( \mu F \) version of these operators. The compilation of `infer` passes the distribution over states to the \( \mu F \) version of `infer`. The inference is thus aware of the distribution over states at the previous step.

\[
C(\text{infer } e) = \text{fun } \sigma -> \\
\text{let } \mu u, \sigma' = \text{infer}(C(e), \sigma) \text{ in } (\mu u, \sigma')
\]

The compilation of a node declaration generates two definitions: the transition function `f_step` and the initial state `f_init`. The transition function is the result of compiling the body of the node with an additional argument to capture the input. The initial step is generated by the allocation function \( \mathcal{A} \) which follows the definition of the initial state in the semantics of Section 3.3 (see the Appendix C for the complete definition).

\[
C(\text{let node } f x = e) = \\
\text{let } f\_\text{step} = \text{fun } (s,x) -> C(e)(s) \\
\text{let } f\_\text{init} = \mathcal{A}(e)
\]

### 4.3 Semantics Equivalence

We showed how to compile ProbZelus to \( \mu F \) a simple functional language with no loops, no recursion, and no higher-order functions, extended with the probabilistic operators. This language is similar to the kernel of \( \mu F \) for which a measure-based probabilistic semantics is defined (see also Figure 16 of Appendix B.2).
We can now prove that the semantics of the generated code corresponds to the semantics of the source language described in Section 3.3.

**Theorem.** For all ProbZelus expression \( e \), for all state \( s \) and environment \( y \):

- if \( \text{kindOf}(e) = \text{D} \) then \( \{ e \}_y^\pi(s) = \{ C(e) \}_y(s), \) and,
- if \( \text{kindOf}(e) = \text{P} \) then \( \{ e \}_y^\pi(s) = \{ C(e) \}_y(s). \)

**Proof.** The proof is done by induction on the structure of \( e \). As an example consider the expression \( \text{sample}(e) \). If this expression is well-typed and since typing is preserved by compilation (see Lemma C.1) \( \text{kindOf}(C(e)) = \text{D} \). Using the induction hypothesis on \( \{ C(e) \}_y(s) = \{ e \}_y^\pi(s) \) we have:

\[
\{ C(\text{sample}(e)) \}_y(s) = \\begin{cases} 
\text{fun } s \rightarrow \text{let } \mu, s' = C(e)(s) \text{ in } \\
\text{let } v = \text{sample}(\mu) \text{ in } (v, s') \end{cases}
\]

\[
\lambda U. \int_{\{ \text{sample}(\mu) \}_y(v, s')} \{ \text{let } v = \text{sample}(\mu) \text{ in } (v, s') \}_y(\mu/\mu, s'/s')(U)
\]

\[
\lambda U. \text{let } (\mu, s) = C(e)(s) \text{ in } \\
\{ \text{let } v = \text{sample}(\mu) \text{ in } (v, s') \}_y(\mu/\mu, s'/s')(U)
\]

\[
\lambda U. \text{let } (\mu, s') = \{ e \}_y^\pi(s) \text{ in } \int \mu(dv)\delta_{v/s'}(U)
\]

\[
\{ \text{sample}(e) \}_y^\pi(s)
\]

\( \square \)

**Remark.** The probabilistic semantics of \( \mu F \) is commutative (see [35, Theorem 4]). We can thus show that the semantics of a ProbZelus program does not depend on the schedule used by the compiler to order the equations of local definitions.

## 5 Inference

The measure-based semantics of \( \text{infer} \) presented in Section 3.3 and Section 4.3 includes often intractable integrals. An additional challenge is to design inference techniques that can operate in bounded memory to be practical in a reactive context where the inference is a non-terminating process.

In this section, we show how to adapt particle filtering [15] to explicitly handle the state of the transition functions. We then present a novel implementation of delayed sampling, a recently proposed semi-symbolic inference, which enables partial exact inference over infinite streams in bounded memory for a large class of models including state-space models like the robot example in Figure 1 in Section 2.

### 5.1 Particle Filtering

In conventional probabilistic programming, the operational interpretation of a model is an importance sampler that randomly generates a sample of the model together with an importance weight measuring the quality of the sample.

Following the conventions of Section 3.3 we write \( \{ e \}_y^\pi \) for the semantics of a deterministic expression, and \( \{ e \}_y^{\pi,w} \) for the semantics of a probabilistic expression. The additional argument \( w \) captures the weight. The probabilistic operator \( \text{sample} \) draws a sample from a distribution without changing the score. \( \text{observe} \) increments the score by the likelihood of the observation. A deterministic expression can be lifted in a probabilistic context: the corresponding sample is the return value of the expression and the score is unchanged. The \text{let} \ construct illustrates that the score is accumulated along the execution path (the complete semantics is presented in Figure 19 of Appendix D).

\[
\{ \text{sample}(e) \}_y^{\pi,w} = (\text{draw}(\{ e \}_y^\pi), w)
\]

\[
\{ \text{observe}(e_1, e_2) \}_y^{\pi,w} = (\mu = \{ e_1 \}_y^{\pi,w} \text{ in } (\bigcup_i w_i \cdot \mu_{\text{pdf}}(\{ e_2 \}_y^{\pi,w}))
\]

\[
\{ e \}_y^{\pi,w} = \{ e \}_y^{\pi}, w \text{ if } \text{kindOf}(e) = \text{D}
\]

\[
\{ \text{let } p = e_1 \text{ in } e_2 \}_y^{\pi,w} = \{ \text{let } v_1, w_1 = \{ e_1 \}_y^{\pi,w} \text{ in } \{ e_2 \}_y^{\pi,w}(v_1/p) \}
\]

Such a sampler is the basis of a particle filter or a bootstrap filter [15]. \( \text{infer} \) independently launches \( N \) particles. At each step, each particle samples the distribution of states \( \sigma \) obtained at the previous step and executes the sampler to compute a pair (result, state) along with its weight. The resulting pairs are then normalized according to their weights to form a categorical distribution \( \mu \) over pairs of values and states (we write \( w_i = w_i / \sum_{i=1}^{N} w_i \) for the normalized weights). This distribution is then split into the distribution of returned values and the distribution of next states.

\[
\{ \text{infer}(\text{fun } s \rightarrow e, \sigma) \}_y = \\
\text{let } \mu = \lambda U. \sum_{i=1}^{N} \text{let } s_i = \text{draw}(\{ \sigma \}_y) \text{ in } \\
\{ \text{fun } s \rightarrow e \}_y^{\pi,w}(s_i) \text{ in } \text{let } (\pi_1(\mu), \pi_2(\mu))
\]

**Remark.** The resampling step requires the ability to clone particles in the middle of the execution. A classic technique is to compile the model in continuation passing style (CPS) [33] and use the probabilistic constructs \( \text{sample} \) and \( \text{observe} \) as checkpoints for resampling. In our context, the compilation presented in Section 4 externalizes the state of the transition functions. Duplicating the state effectively clones a particle during its execution. The code does not need to be compiled in CPS form and we avoid the alignment problem [24].

### 5.2 Delayed Sampling

The basis of our streaming inference algorithm is delayed sampling, which we first review to explain its conceptual approach. Delayed sampling is an inference technique combining partial exact inference with approximate particle filtering to reduce estimation errors [23, 27].
The probabilistic sample(e) adds a new random variable to the graph without drawing a sample. observe(e₁, e₂) adds a new random variable \( X \sim \mu \) where \( \mu \) is defined by \( e₁ \), then computes a concrete value \( v \) for \( e₂ \) and conditions the graph by observing the value \( v \) for \( X \). As for the particle filter, the score is incremented by the likelihood of the observation.

**Symbolic Computations.** The functions value, assume, and observe used in Figure 8 rely on the following mutually recursive lower level operations (\( Y \) is the parent of \( X \)) [27]:

\[
X, g' = initialize(\mu, Y, g) \quad \text{add a new node} \ X \ \text{with a distribution} \ p_{X|Y} = \mu \ \text{as a child of} \ Y \ in \ g.
\]

\[
g' = marginalize(X, g) \ \text{compute and store} \ p_X \ \text{in} \ g' \ \text{from} \ p_Y \ \text{and} \ p_{X|Y} \ \text{where} \ p_Y \ \text{and} \ p_{X|Y} \ \text{are in} \ g.
\]

\[
g' = realize(X, v, g) \ \text{assign in} \ g' \ \text{a concrete value to a random variable} \ X.
\]

\[
g' = condition(Y, g) \ \text{compute} \ p_{Y|X} \ \text{given} \ p_X, p_{X|Y}, \ \text{and a concrete value} \ X = v \ \text{where} \ v \ \text{is in} \ g.
\]

In the class of Bayesian networks maintained by the delayed sampler, marginalization w.r.t. a parent node, and conditioning a parent on the value of a child are tractable operations.

To reflect these operations, nodes are characterized by a state (see Figures 4 and 9). **Initialized nodes** are random variables with a conditional distribution \( p_{X|Y} \) where the parent \( Y \) has no concrete value yet. **Marginalized nodes** are random variables with a marginal distribution \( p_X \) that incorporate the distributions of the ancestors. **Realized nodes** are random variables that have been assigned a concrete value via sampling or observation.

The evaluation function value(e, g) forces the realization by sampling of all the random variables referenced in \( e \) to produce a concrete value. Similarly, the function observe(X, v, g) realizes a variable \( X \) with a given observation \( v \). The realization of a random variable comprises three steps: (1) compute the distribution \( p(x) \) by recursively marginalizing the parents from a root node, (2) sample a value, or use the observation, and (3) use the concrete value to update the children and condition the parent which removes the dependencies.

The function assume(\( \mu, g \)) adds a new node to the graph and is defined case by case on the shape of the symbolic term \( \mu \). If there is a conjugacy relationship between \( \mu \) and a random variable \( Y \) present in the graph, e.g., \( \mu = Bernoulli(Y) \) with \( Y \sim Beta(\alpha, \beta) \), a new initialized node \( X \sim \mu \) is added as a child of \( Y \). Otherwise, since symbolic computation is not possible, dependencies are broken by realizing the random variables that appear in \( \mu \), e.g., \( \mu' = Bernoulli(value(Y, g)) \), and \( X \sim \mu' \) is added as a new root node.

**Inference.** The inference scheme is similar to the particle filter. At each step, the inference draws \( N \) states from \( \sigma \) to execute the transition function. For each particle, execution starts with the graph computed at the previous step and returns a pair of symbolic terms (result, state), the particle

\[
\langle \text{op}(e), \mu, v, w \rangle = \langle \langle e \rangle, \langle e \rangle, \mu, v, w \rangle \ \text{in} \ (\text{app}(\text{op}(e'), g_e, w_e))
\]

\[
\text{if} \ e \ \text{then} \ e_1 \ \text{else} \ e_2 \ \text{if} \ e \ \text{then} \ e_1 \ \text{else} \ e_2 \ \text{let} \ e', g_e, \text{we} = \langle e \rangle \ \text{in} \ \langle e \rangle \ \text{in} \ \langle X, g', w' \rangle
\]

\[
\langle \text{sample}(e), \mu, v, w \rangle = \langle e, g_e, \text{we} = \langle e \rangle \ \text{in} \ \langle X, g', w' \rangle
\]

\[
\langle \text{observe}(e_1, e_2), \mu, v, w \rangle = \langle e_1, g_e, \text{we} = \langle e \rangle \ \text{in} \ \langle X, g', w' \rangle = \langle e \rangle \ \text{in} \ \langle X, g', w' \rangle
\]

Figure 8. Delayed sampling sampler. Expressions return a pair (symbolic expression, weight). In addition to the importance weight, each particle exploits conjugacy relationships between pairs of random variables to maintains a graph: a Bayesian network representing closed-form distributions involving subsets of random variables. Observations are incorporated by analytically conditioning the network. Particles are thus only required to draw sample when forced to, i.e., when exact computation is not possible, or when a concrete value is required.

To perform analytic computations, delayed sampling manipulates symbolic terms where random variables are referenced in the graph. The semantics of an expression \( \langle e \rangle \ \text{in} \ \langle Y, g, w \rangle \) takes an additional argument \( g \) for the graph and returns a symbolic term, an updated weight, and an updated graph. Given a graph, a symbolic term can be evaluated into a concrete value by sampling the random variables that appear in the term. The graph can be accessed and modified using the three following functions defined in [27].

\[
v, g' = value(e, g) \ \text{evaluate a symbolic term and return a concrete value.}
\]

\[
X, g' = assume(\mu, g) \ \text{add a random variable} \ X \sim \mu \ \text{to the graph and return the variable.}
\]

\[
g' = observe(X, v, g) \ \text{condition the graph by observing the value} \ v \ \text{for the variable} \ X.
\]

\[
\text{Inference.} \ \text{The inference scheme is similar to the particle filter. At each step, the inference draws} \ N \ \text{states from} \ \sigma \ \text{to execute the transition function. For each particle, execution starts with the graph computed at the previous step and returns a pair of symbolic terms (result, state), the particle}
\]

\[
\text{The probabilistic sample(e) adds a new random variable to the graph without drawing a sample. observe(e₁, e₂) adds a new random variable} \ X \sim \mu \ \text{where} \ \mu \ \text{is defined by} \ e₁, \ \text{then computes a concrete value} \ v \ \text{for} \ e₂ \ \text{and conditions the graph by observing the value} \ v \ \text{for} \ X. \ \text{As for the particle filter, the score is incremented by the likelihood of the observation.}
\]

\[
\text{Symbolic Computations.} \ \text{The functions value, assume, and observe used in Figure 8 rely on the following mutually recursive lower level operations (Y is the parent of X) [27]:}
\]

\[
X, g' = initialize(\mu, Y, g) \ \text{add a new node} \ X \ \text{with a distribution} \ p_{X|Y} = \mu \ \text{as a child of} \ Y \ \text{in} \ g.
\]

\[
g' = marginalize(X, g) \ \text{compute and store} \ p_X \ \text{in} \ g' \ \text{from} \ p_Y \ \text{and} \ p_{X|Y} \ \text{where} \ p_Y \ \text{and} \ p_{X|Y} \ \text{are in} \ g.
\]

\[
g' = realize(X, v, g) \ \text{assign in} \ g' \ \text{a concrete value to a random variable} \ X.
\]

\[
g' = condition(Y, g) \ \text{compute} \ p_{Y|X} \ \text{given} \ p_X, p_{X|Y}, \ \text{and a concrete value} \ X = v \ \text{where} \ v \ \text{is in} \ g.
\]

In the class of Bayesian networks maintained by the delayed sampler, marginalization w.r.t. a parent node, and conditioning a parent on the value of a child are tractable operations.

To reflect these operations, nodes are characterized by a state (see Figures 4 and 9). **Initialized nodes** are random variables with a conditional distribution \( p_{X|Y} \) where the parent \( Y \) has no concrete value yet. **Marginalized nodes** are random variables with a marginal distribution \( p_X \) that incorporate the distributions of the ancestors. **Realized nodes** are random variables that have been assigned a concrete value via sampling or observation.

The evaluation function value(e, g) forces the realization by sampling of all the random variables referenced in \( e \) to produce a concrete value. Similarly, the function observe(X, v, g) realizes a variable \( X \) with a given observation \( v \). The realization of a random variable comprises three steps: (1) compute the distribution \( p(x) \) by recursively marginalizing the parents from a root node, (2) sample a value, or use the observation, and (3) use the concrete value to update the children and condition the parent which removes the dependencies.

The function assume(\( \mu, g \)) adds a new node to the graph and is defined case by case on the shape of the symbolic term \( \mu \). If there is a conjugacy relationship between \( \mu \) and a random variable \( Y \) present in the graph, e.g., \( \mu = Bernoulli(Y) \) with \( Y \sim Beta(\alpha, \beta) \), a new initialized node \( X \sim \mu \) is added as a child of \( Y \). Otherwise, since symbolic computation is not possible, dependencies are broken by realizing the random variables that appear in \( \mu \), e.g., \( \mu' = Bernoulli(value(Y, g)) \), and \( X \sim \mu' \) is added as a new root node.
when a node is realized. All nodes that have been neither sampled nor observed are thus kept in the graph even if they are no longer referenced by the program. In a reactive programming context, such an implementation can consume unbounded memory.

**Bounded Delayed Sampling.** A simple mitigation is to limit the scope of symbolic computations to one time step and discard the graph at the end of each time step. We call this inference technique **bounded delayed sampling** (BDS).

BDS performs symbolic computations during the execution of a time step, and whenever possible, delays the sampling until the end of the instant. Like the particle filter, BDS guarantees a bounded-memory execution. For each particle, the size of the graph is bounded by the number of variables introduced during a time step, which by construction, is bounded for any valid ProbZelus program.

**Streaming Delayed Sampling.** Compared to the original delayed sampling algorithm, BDS loses the ability to perform symbolic computations using variables defined at different time steps. This can result in a significant loss of precision for models with inter-steps dependencies such as the robot example of Figure 1. To adapt delayed sampling to streaming settings while keeping its maximum accuracy, we designed a delayed sampler that is **pointer-minimal** where nodes that are no longer referenced by the program can be eventually removed. We call this inference **streaming delayed sampling** (SDS). SDS enables partial exact inference in bounded memory for a large class of models.

In the original implementation of delayed sampling, graph nodes need to access their parents and children. Marginalization requires access to the parent to incorporate the ancestor distribution. Realization requires access to both the parent and the children of a node to update their respective distributions with the concrete value assigned to the node.

In the pointer minimal implementation, initialized nodes only keep a pointer to their parent to follow the ancestor chain during marginalization and marginalized nodes only keep a pointer to a marginalized child (Delayed Sampling imposes that a node always has at most one marginalized child). Compared to the original implementation, marginalized nodes only keep track of one child, and marginalization turns backward pointers to the parent node into forward pointers to the marginalized child. Note that this implementation prevents updating the children when the parent is realized, and prevents conditioning a parent when a child is realized. Instead, when marginalizing a node, the sampler first checks if the parent is realized to apply the update. Symmetrically, to realize a node, the sampler first checks if the children are realized and, if necessary, conditions the distribution before assigning the concrete value.

Figure 9 shows the evolution of the graph during one step for the robot example of Figure 1. At the end of the step, the value of \texttt{pre }x \texttt{is updated. The previous value is not referenced anymore by the program and the node can be removed from the graph. In the original implementation, backward pointers between marginalized nodes prevent the collection (see Figure 4).}

**Limitations.** With SDS, models like the robot example that only maintain bounded chains of dependencies between variables are guaranteed to be executed in bounded memory. The class of models that can be executed in bounded memory...
with our pointer-minimal implementation thus already comprises state-space models like Kalman filters, and models for learning unknown constant parameters from a series of observations (e.g., computing the bias of a coin from a succession of flips) where variables introduced at each step are immediately realized.

However, unbounded chains can still be formed if the program keeps a reference to a constant variable that is never realized. In the following example, at each step, a new variable \( x \) is added as a child of \( \text{pre} \ x \) and then marginalized for the observation. But \( p1 \) keeps a reference to the initial variable \( i \) which is never realized and thus forms an unbounded chain between \( i \) and \( x \).

```plaintext
let proba p1 (xo, obs) = (i, o) where
  rec init i = sample (gaussian (xo, 1.))
  and x = sample (gaussian (i -> pre x, 1.))
  and () = observe (gaussian (x, 1.), obs)
```

In addition, in ProbZelus, at each step the inference returns a snapshot of the current distribution without forcing the realization of any node in the graph. Compared to the original delayed sampling implementation, initialized nodes can be inspected without being realized. It is thus possible to form unbounded chains of initialized nodes which cannot be pruned even when nodes are no longer referenced in the program due to the backward pointers to the parent in initialized nodes. In the following example, at each step, the variable \( x \) is added as a child of \( \text{pre} \ x \), but without observation these variables remain initialized for ever.

```plaintext
let proba p2 (xo) = x where
  rec x = sample (gaussian (xo -> pre x, 1.))
```

To mitigate these issues, we can force the realization of trailing nodes at each step as in \textit{bounded delayed sampling} or use a sliding window. Alternatively, the \texttt{value} (eval) function is available to the programmer and can be used to implement any strategy to force the evaluation of the nodes. For instance, the previous example can be adapted to execute in bounded memory:

```plaintext
let proba p2’ (xo) = x where
  rec x = sample (gaussian (xo -> pre x, 1.))
  and _ = eval (xo -> pre x)
```

### 6 Evaluation

We next evaluate the performance of ProbZelus on a set of benchmarks that illustrate multiple aspects of the language: inferring fixed parameters from observations, online trajectory estimation, inference-in-the-loop. For these examples, we compare the accuracy and the latency cost of the three inference techniques: PF, BDS, and SDS. Appendix E details our implementation as an extension of the Zelus compiler.

#### Benchmarks

The models used in the experiments are summarized in Table 1 (a detailed description along with the code of the benchmarks is given in Appendix F.1). Two models infer fixed parameters from a series of observations. Beta-Bernoulli estimates the parameter of a Bernoulli distribution from a series of binary observations (e.g., the bias of a coin). Gaussian-Gaussian estimates the mean and variance of a Gaussian distribution from a series of observations. The accuracy metric is the Mean Squared Error (MSE) of the inferred parameters compared to their exact values.

Two models infer the state of a moving agent from noisy observations. Kalman-1D is a one-dimensional Kalman filter that models an agent that estimates its trajectory from noisy observations. Outlier adapted from \cite{16} models the same situation as Kalman-1D, but the sensor occasionally produces invalid readings. This model infers both the trajectory of the agent, and the bias of the sensor. The accuracy metric is the Mean Squared Error (MSE) of the inferred trajectory compared to the exact positions.

Two models use inference-in-the-loop (IITL). Robot is the robot example of Section 2 and the accuracy metric is the LQR loss. SLAM (Simultaneous Localization and Mapping) adapted from \cite{15} models an agent that estimates its position and a map of its environment. In this simplified version, the agent moves in a one-dimensional grid where each cell is either black or white. The robot’s wheels may slip causing the robot to unknowingly stay in place (noisy motion), and the sensor is not perfect and may accidentally report the wrong color (noisy observations). At each step the robot uses the inferred position to decide its next move. The accuracy metric is the MSE of both the position and the map.

MTT (Multi-Target Tracker) adapted from \cite{26} is a model where there are a variable number of targets with linear-Gaussian motion models with a state space of 2D position and velocity, producing linear-Gaussian measurements of the position at each time step. Targets randomly appear according to a Poisson process and each disappear with fixed probability at each step. Measurements do not identify which target they came from, and “clutter” measurements that come not from targets but from some underlying distribution add

<table>
<thead>
<tr>
<th>Fixed</th>
<th>Moving</th>
<th>IITL</th>
<th>Metric</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beta-Bernoulli</td>
<td>✓</td>
<td></td>
<td>MSE</td>
</tr>
<tr>
<td>Gaussian-Gaussian</td>
<td>✓</td>
<td></td>
<td>MSE</td>
</tr>
<tr>
<td>Kalman-1D</td>
<td>✓</td>
<td>✓</td>
<td>MSE</td>
</tr>
<tr>
<td>Outlier</td>
<td>✓</td>
<td>✓</td>
<td>MSE</td>
</tr>
<tr>
<td>Robot</td>
<td>✓</td>
<td>✓</td>
<td>LQR</td>
</tr>
<tr>
<td>SLAM</td>
<td>✓</td>
<td>✓</td>
<td>MSE</td>
</tr>
<tr>
<td>MTT</td>
<td>✓</td>
<td></td>
<td>MOTA*</td>
</tr>
</tbody>
</table>

Table 1. Benchmarks with: inference of \textit{fixed} parameters from observations, estimation of a \textit{moving} state (state-space model), and \textit{inference-in-the-loop} (IITL).
to observations, complicating inference of which measurements are associated to which targets. The accuracy metric is expected MOTA+ = (1/MOTA) − 1 where MOTA ∈ [0, 1] is the Multiple Object Tracking Accuracy [3].

**Experimental Setup.** All the experiments were run on a server with 32 CPUs (2.60 GHz) and 128 GB memory. We ran all the benchmarks for 500 steps. In all cases, the inference runs in bounded memory (see Appendix F.3).

For each algorithm, we evaluated how much time it required to achieve 90% of runs close to a loss target (out of 1000 runs total):

$$|\log(P_{90\%}(\text{loss})) − \log(\text{loss}_{\text{target}})| < 0.5.$$

For each benchmark, the baseline is the median loss of SDS at 1000 particles as loss_{\text{target}} for that benchmark. We measured the number of particles required to achieve this loss, and then measured the total execution time at this particle count for 500 steps (in Appendix F.2 we also evaluate loss and step latency across a fixed range of particle counts).

**Results.** Figure 10 shows the results. The height of each bar is the median total execution time, and the error bars are 90% and 10% quantiles, aggregated over 1000 runs. Each bar is labeled with the minimum number of particles required to achieve the accuracy threshold, accurate to 1.5 significant digits (100, 150, 200, 250, …). We observe that SDS is able to compute an exact solution for Beta-Bernoulli, Kalman-1D, and Robot. In all these examples 1 particle is already enough to reach the target accuracy. Overall, the results show that the number of particles required to reach the desired accuracy with PF implies a significant slowdown compared to SDS. Moreover, the SLAM and MTT benchmarks show that, in some cases, PF is not an option: the target accuracy was not reached with 15,000 and 2,500 particles, respectively, at which point PF was already 10 times slower than SDS and we stopped the experiments.

As expected, BDS performance numbers are between those of PF and SDS. At worst, when there is no possible intra-step symbolic computations (e.g., Beta-Bernoulli), BDS behaves like a particle filter and requires as many particles as PF. At best, BDS performs as well as SDS (e.g., Outlier).

Additionally, Figure 10 also shows that for a given number of particles, the overhead induced by managing the delayed sampling graph is significant. Compared to BDS and SDS, depending on the benchmark, it is possible to use 2 to 4 times as many particles for PF with the same execution time. However, this is not enough to match the gain in accuracy.

**Alternative Baselines.** The results presented in Figure 10 do not quantify the speedup of SDS on the SLAM and MTT benchmarks because the other inference algorithms time out. To evaluate speedups on these two benchmarks, we used PF as an alternative baseline instead of SDS. Figure 11 presents the execution time of PF, BDS, and SDS to reach a loss close to the median of PF with 2000 and 4000 particles.

We observe that SDS requires a much smaller number of particles to reach similar accuracy which translates into speedups ranging from 10^3 (MTT-2000) to 10^4 (SLAM-4000). BDS requires either a similar or smaller numbers of particles. But the overhead introduced by the graph manipulations mostly translates in slowdowns compared to PF.

**7 Related Work**

**Probabilistic Programming.** Over the last few years there has been a growing interest on probabilistic programming languages. Some languages like BUGS [25], Stan [9], or Augur [21] offer optimized inference technique for a constrained subset of models. Other languages like WebPPL [18], Edward [37], Pyro [6], or Birch [26] focus on expressivity allowing the specification of arbitrary complex models. Compared to these languages, ProbZelus can be used to program reactive models that typically do not terminate, and inference can be run in parallel with deterministic components that interact with an environment.
A notion of processes that can be composed in parallel, again, their interfaces only support inference on a complete probabilistic model of their environment (e.g., object tracking).

These models cannot be expressed in ProbZelus which relies on a probabilistic model. Modeling uncertainty is a primary element of control systems, respectively.

Reactive Languages with Uncertainty. Lutin is a language for describing non-deterministic reactive systems for testing and simulation [32], while but Lutin supports weighted sampling to describe constrained random scenarios, it does not support inference. ProPL [30] is a language to describe probabilistic models for process that evolve over a period of time. This language also extends a probabilistic language with a notion of processes that can be composed in parallel, but compared to ProbZelus, ProPL focuses on a constrained class of models that can be interpreted as Dynamic Bayesian Networks (DBN), and relies on standard DBN inference techniques. In the same vein, CTPPL [31] is a language to describe continuous-time processes where the amount of time taken by a sub-process can be specified by a probabilistic model. These models cannot be expressed in ProbZelus which relies on the synchronous model of computation. It would be interesting to investigate how to extend ProbZelus to continuous-time models based on Zelus’ support for ordinary differential equations (ODE) [7].

Inference. Researchers have proposed streaming inference algorithms, including variational [8], or sampling-based [16, 19] approaches. Popular languages like Stan, Edward, or Pyro, offer support to stream data through the model during inference to handle large datasets. However, compared to ProbZelus, the model must be defined a priori and does not evolve during the inference.

The Anglican and Birch probabilistic programming languages support delayed sampling [27]. These languages do not support streaming inference or reactive programming. Again, their interfaces only support inference on a complete probabilistic model.

8 Conclusion

Modeling uncertainty is a primary element of control systems for tasks that operate under the assumption of a probabilistic model of their environment (e.g., object tracking). While synchronous languages have developed as a prominent way to develop control applications, to date there has been limited work in these languages on programming language support for modeling uncertainty.

In this paper we present ProbZelus, the first synchronous probabilistic programming language that lifts emerging abstractions for probabilistic programming into the reactive setting thus enabling inference-in-the-loop. Moreover, our streaming delayed sampling algorithm provides efficient semi-symbolic inference while still satisfying a key requirement of control applications in that they must execute with bounded resources.

Our results demonstrate that ProbZelus enables us to write, in the very same source, a deterministic model for the control software and a probabilistic model for its behavior and environment with complex interactions between the two.

Acknowledgments

This work was supported in part by the MIT-IBM Watson AI Lab and the Office of Naval Research (ONR N00014-17-1-2699).

References


A ProbZelus

In this section, we provide the complete definitions of the ProbZelus type system and semantics for the kernel language introduced Section 3 extended with the probabilistic operator factor(e) which is equivalent to observe(exp(1), e). Intuitively, factor can directly update the weight of the execution path with the value of an expression e.

A.1 Typing

The type system that discriminates deterministic from probabilistic expressions is defined Figure 12. To simplify the presentation, we ignored datatypes polymorphism.

The sub-typing rule indicates that any deterministic expression can be lifted into a probabilistic one. Expressions like constants, variables, and last are deterministic. The kind of classic Zelus expressions (pairs, ap, local definitions, present, and reset) is the kind of their body. Similarly, the kind of equations is the kind of their defining expression, and parallel composition imposes the same kind for all the equations. Note that it is always possible to compose deterministic and probabilistic computations. For rules where all sub-expressions share the same kind k we enforce the use of the sub-typing rule to lift deterministic expressions.

The expressions sample, factor, and observe are probabilistic. The transition from probabilistic to deterministic is realized via infer: a deterministic expression whose body is always predictable. Probabilistic expressions can thus only occur under an infer.

Other Static Analyses. The Zelus compiler statically checks initialization, and causality of the program [7]. These two analyses guarantee that there exists a schedule of parallel equations that makes the streams productive. Extending these analyses to the probabilistic operators is straightforward: probabilistic operators can be treated as external operators.

A.2 Co-iterative Semantics

The co-iterative semantics of ProbZelus’s deterministic processes is inspired by [12] and defined Figure 13.

A node is a stream function of type T → T’. In addition to the state, the transition function thus takes an additional input of type T and returns a pair (result, next state)

\[ \text{CoNode}(T, T’, S) = S \times (S \rightarrow T \rightarrow T’ \times S). \]

The transition function of a variable always returns the corresponding value stored in the environment y. The semantics of last x is a simple access to a special variable x_last. present e -> e1 else e2 introduced in Section 2 returns the value of e1 when e is true and the value of e2 otherwise. The state (s, s1, s2) stores the state of the three sub-expressions. The transition function lazily executes e1 or e2 depending on the value of e and returns the updated state.

The state of a set of scheduled locally recursive definitions e where rec E comprises three parts: the value of the local variables at the previous step which can be accessed via the last operator, the state of the defining expressions, and the state of expression e. The initialization stores the initial values introduced by init and the initial states of all sub-expressions. The transition function incrementally builds the local environment defined by E. First the environment is populated with a set of fresh variables x1_last initialized with the values stored in the state that can then be accessed via the last operator. Then the environment is extended with the definition of all the variables yk by executing all defining expressions (where \( \{x_1\}_{1..k} \cap \{y_1\}_{1..n} = \{x_1\}_{1..k} \)). Finally, the expression e is executed in the final environment. The updated state contains the value of the initialized variables defined in E that will be used to start the next step, and the updated state of the sub-expressions.

Probabilistic Extensions. The semantics of the probabilistic part of ProbZelus, defined Figure 14, follows the same structure as the deterministic semantics but defines measures over all possible executions as in [39]. In particular a succession of computation is interpreted as sequentially integrating over the results of the preceding computations.

As for deterministic nodes, the transition function of a probabilistic node of type T → T’ takes an additional argument and returns a measure over pairs (result, next state).

\[ \text{CoPNode}(T, T’, S) = S \times (S \rightarrow T \rightarrow (\Sigma_{T \times S} \rightarrow [0, \infty])) \]

A.3 Alternative semantics

We could give different semantics to ProbZelus. For example, consider the following probabilistic node.

\[ \text{let proba kahn_vs_scott () = p where} \]
\[ \text{rec init p = sample(beta(1, 1))} \]
\[ \text{and () = observe(bernoulli(p), true)} \]

With the semantics defined Section 3, this program produces the stream of distribution: \( \text{Beta}(2, 1), \text{Beta}(3, 1), \ldots \). Note that, even though, p is defined as a constant, its distribution evolves at each steps.

Since the observe statement uses the constant true, we know that p is necessarily 1. An alternative semantics could thus returns the constant stream of distributions: \( \delta_1 \).

B The \( \mu F \) language

Similarly to ProbZelus, we extend \( \mu F \) with the probabilistic operator factor. We now present the complete type system and semantics for \( \mu F \).

B.1 Typing

The type system defined Figure 15 is similar to the one Figure 12 to distinguish deterministic from probabilistic expressions, but with additional restrictions since the compiled code is in a more constrained form. Whenever possible
we require sub-expressions to be deterministic, that is, in pairs, operator applications (including `sample`, `factor`, and `observe`), function calls, and the condition of a `if/then/else`. These restrictions simplify the presentation of the semantics but do not reduce the expressiveness of the language since it is always possible to introduce additional local definitions to name intermediate probabilistic expressions. For example `if sample(bernoulli(0.5)) then ... can be rewritten

\begin{verbatim}
let b = sample(bernoulli(0.5)) in if b then ...
\end{verbatim}

\section{B.2 Semantics of $\mu F$}

The semantics of $\mu F$ follows [39]. In a deterministic context `kindOf(e) = D`, the semantics $\llbracket e \rrbracket_D$ of an expression is the classic interpretation of a strict functional language. In a probabilistic context (`kindOf(e) = P`), we define a the measure-based semantics $\llbracket e \rrbracket_P$.

The probabilistic semantics of $\mu F$ is presented in Figure 16. A deterministic expression is lifted to a probabilistic expression using the Dirac delta measure applied to the value of the expression computed by the deterministic semantics. As in Section 3.3, a local definition `let x = e1 in e2` is interpreted as integrating $e_2$ over the measure defined by $e_1$. The semantics of the probabilistic operators is the following: `sample(e)` returns the distribution $\llbracket e \rrbracket_P$. `factor(e)` returns a measure defined on the singleton space $\{\}$ whose value is $\exp(\llbracket e \rrbracket_P)$. `observe(e1, e2)` is similar but the score is the density function of the distribution $\llbracket e_1 \rrbracket_P$ applied to $\llbracket e_2 \rrbracket_P$.

\section{C Compilation}

Figure 18 presents the entire compilation function from ProbZelus to $\mu F$ introduced Section 4. Figure 17 presents the allocation function.

\textbf{Lemma C.1.} The compilation preserves the kind (deterministic or probabilistic) of the expressions. For any expression $e$, if $G \vdash^k e : t$, there exists $G'$ and $t'$ such that $G' \vdash^k C(e) : t'$.

\textbf{Proof.} By induction on the structure of $e$. \hfill $\square$

\textbf{Remark.} The compilation presented in Figure 18 generates a function for each sub-expression. However, in most cases it is possible to simplify the code using static reduction. For instance, a constant can directly be compiled into a constant.
\[
\begin{align*}
[c]_y^e &= (c) \\
[e]_y^e &= \lambda s. (c, s) \\
x_1^e &= () \\
x_i^e &= \lambda s. (y(x), s) \\
\text{last } x^e_y &= () \\
\text{last } x_i^e_y &= \lambda s. (y(x \text{ last}), s) \\
[e_1 \cdot e_2]^y &= ([e_1]^y \cdot [e_2]^y) \\
\text{let } v_1, s'_1 = [e_1]^y (s_1) \text{ in} \\
&\quad \text{let } v_2, s'_2 = [e_2]^y (s_2) \text{ in } ((v_1, v_2), (s'_1, s'_2)) \\
\text{present } e \rightarrow e_1 \text{ else } e_2^y &= ([e_1]^y, [e_1]^y, [e_2]^y) \\
\text{present } e \rightarrow e_1 \text{ else } e_2^y &= \lambda s_1, s_2. \text{let } v, s' = [e_1]^y (s) \text{ in} \\
&\quad \text{if } v \text{ then let } v_1, s'_1 = [e_1]^y (s_1) \text{ in } (v_1, (s'_1, s'_2)) \\
&\quad \text{else let } v_2, s'_2 = [e_2]^y (s_2) \text{ in } (v_2, (s'_1, s'_2)) \\
\text{reset } e_1 \text{ every } e_2^y &= ([e_1]^y \cdot [e_1]^y \cdot [e_2]^y) \\
\text{reset } e_1 \text{ every } e_2^y &= \lambda s_0, s_1, s_2. \text{let } v_2, s'_2 = [e_2]^y (s_2) \text{ in} \\
&\quad \text{let } v_1, s'_1 = [e_1]^y (if \ v_2 \text{ then } s_0 \text{ else } s_1) \text{ in } (v_1, (s_0, s'_1, s'_2)) \\
\text{e where} \\
\text{rec } \text{init } x_1 = c_1 \text{ and } \ldots \\
\text{and init } x_k = c_k \\
\text{and } y_1 = e_1 \text{ and } \ldots \\
\text{and } y_n = e_n \\
&\mapsto \left(\begin{array}{c}
\lambda((m_1, \ldots, m_k), (s_1, \ldots, s_n), s) \\
\text{let } y_1 = y[m_1/x_1, \text{ last}] \text{ in } \\
\left(\begin{array}{c}
\lambda v. \lambda s. [e_1]^y[v/v_1, y_1] \text{ in } \\
\text{let } v_k = y_{k-1}[m_k/x_k, \text{ last}] \text{ in } \\
\text{let } v_1, s'_1 = [e_1]^y[v_1, y_1] \text{ in } \\
\text{let } y'_n = y_{n-1} [v_n/y_n] \text{ in } \\
\text{let } v, s' = [e_1]^y[v/v_1, y_1] \text{ in } \\
\left(\begin{array}{c}
\lambda(v, ((v_1)^n, \ldots, v_n[x_k]), (s'_1, \ldots, s'_n), s') \\
\end{array}\right)
\end{array}\right)
\right)
\end{align*}
\]

Figure 13. Co-iterative semantics of deterministic ProbZelus programs. For local definitions each initialized variable is defined in a subsequent equation, i.e., \( \{x_i\}_{i=1}^k \cap \{y_j\}_{j=1}^n = \{x_i\}_{i=1}^k \).
\[
\begin{align*}
\{e\}_{\gamma}^i &= \mathbb{P} \text{ if } \text{kindOf}(e) = \text{D} \\
\{e\}_{\gamma}^i &= \lambda s. \lambda U. \mathbb{P}_{\text{exp}}^{\gamma}(s)(U) \text{ if } \text{kindOf}(e) = \text{D} \\
\{(e_1, e_2)\}_{\gamma}^i &= \lambda s_1, s_2. \lambda U. \text{let } \mu_1 = \{e_1\}_{\gamma}^{\gamma}(s_1) \text{ in} \\
&\quad \int \mu_1(dv_1, ds_1') \text{ let } \mu_2 = \{e_2\}_{\gamma}^{\gamma}(s_2) \text{ in} \\
&\quad \int \mu_2(dv_2, ds_2') \delta_{(v_1, v_2, s_1', s_2')}(U) \\
\{\text{init}\}_{\gamma} &= \{(\{e_1\}_{\gamma}, \{e_2\}_{\gamma})\} \\
\{\text{reset}\}_{\gamma} &= \lambda s_0, s_1, s_2. \lambda U. \text{let } \mu_2 = \{e_2\}_{\gamma}^{\gamma}(s_2) \text{ in} \\
&\quad \int \mu(dv_2, ds_2') \text{ let } \mu_1 = \{e_1\}_{\gamma}^{\gamma}(s_1) \text{ in} \\
&\quad \int \mu(dv_1, ds_1') \delta_{(v_1, v_2, s_1', s_2')}(U) \\
\{\text{sample}\}_{\gamma} &= \mathbb{P} \\
\{\text{obs}\}_{\gamma} &= \mathbb{P} \\
\{\text{factor}\}_{\gamma} &= \mathbb{P} \\
\{\text{observe}\}_{\gamma} &= \lambda s_1, s_2. \lambda U. \text{let } \mu, s' = \{e_1\}_{\gamma}^{\gamma}(s_1) \text{ in} \\
&\quad \text{let } v, s' = \{e_2\}_{\gamma}^{\gamma}(s_2) \text{ in} \\
&\quad \mu_{\text{pdf}}(v) \times \delta_{(\iota(s_1', s_2'), (v, s_2'))}(U) \\
\end{align*}
\]

\textbf{Figure 14.} Co-iterative semantics of probabilistic ProbZelas expressions (i.e., \text{kindOf}(e) = \text{P}). For local definitions each initialized variable is defined in a subsequent equation, i.e., \{x_i\}_{i=1..k} \cap \{y_j\}_{j=1..n} = \{x_i\}_{i=1..k}.

\[\]
\[
\begin{align*}
G v^0 & : t & \text{typeOf}(c) = t & G(x) = t & G v^0 e_1 : t & G v^0 e_2 : t & \text{typeOf}(\text{op}) = t_1 \rightarrow^D t_2 & G v^0 e : t_1 \\
G v^p & : t & G v^0 c : t & G v^p x : t & G v^0 (e_1, e_2) : t_1 \times t_2 & G v^p \text{op}(e) : t_2 & \\
G(f) = t_1 \rightarrow^k t_2 & G v^0 e : t_1 & G + \left[t_1/x\right] v^k e_1 : t_2 & G v^0 e_2 : t_1 & G v^p e : t_2 & G v^0 e : t_2 & \\
G v^k f(e) : t_2 & G v^k \text{(fun } x \rightarrow e_1)(e_2) : t & G + \left[t_1/x\right] v^k e : t_2 & G v^p \text{fun } x \rightarrow e_1 \rightarrow^k t_2 & G v^p \text{sample}(e) : t & G v^p e : t_2 & \\
G v^k e_1 : t_1 & G + \left[t_1/x\right] v^k e_2 : t_2 & G v^p \text{let } x = e_1 \text{ in } e_1 : t_2 & G v^p \text{fun } x \rightarrow e_1 \rightarrow^k t_2 & G v^0 e : t_2 & G v^0 e : t_2 & \\
G v^0 \text{observe}(e_1, e_2) : \text{unit} & G v^p \text{factor}(e) : \text{unit} & G v^p \text{infer}(\text{fun } x \rightarrow e_1, e_2) : \text{t dist} & G v^p \text{infer}(\text{fun } x \rightarrow e_1, e_2) : \text{dist} & G v^0 e : \text{t dist} & \\
G v^p \text{let } f = e : G + [t/f] & \\
\end{align*}
\]

Figure 15. Typing of \(\mu F\) with deterministic and probabilistic kinds.

\[
\begin{align*}
\text{let } f = e \rangle_Y &= \gamma[\langle e \rangle_Y/f] \\
\langle d_1, d_2 \rangle_Y &= \text{let } y_1 = \langle d_1 \rangle_Y \text{ in } \langle d_2 \rangle_Y, \\
\langle e \rangle_Y &= \lambda U. \delta_{\text{let}(e)}(U) \text{ if kindOf}(e) = \text{D} \\
\langle e_1(e_2) \rangle_Y &= \lambda U. \langle \langle e_1 \rangle_Y, \langle e_2 \rangle_Y \rangle(U) \\
\text{let } p = e_1 \text{ in } e_2 \rangle_Y &= \lambda U. \int_T \langle e_1 \rangle_Y du \langle e_2 \rangle_Y[U/p](U) \\
\text{if } e \text{ then } e_1 \text{ else } e_2 \rangle_Y &= \lambda U. \text{if } \langle e \rangle_Y \text{ then } \langle e_1 \rangle_Y(U) \text{ else } \langle e_2 \rangle_Y(U) \\
\langle \text{fun } p \rightarrow e \rangle_Y &= \lambda U. \langle e \rangle_Y[U/p] \\
\langle \text{sample}(e) \rangle_Y &= \lambda U. \langle \langle e \rangle_Y \rangle(U) \\
\text{observe}(e_1, e_2) \rangle_Y &= \lambda U. \text{let } \mu = \langle e_1 \rangle_Y \text{ in } \mu_{\text{dist}}(\langle e_2 \rangle_Y) * \delta_{\text{D}}(U) \\
\langle \text{factor}(e) \rangle_Y &= \lambda U. \exp(\langle \langle e \rangle_Y \rangle) * \delta_{\text{D}}(U) \\
\text{infer}(\text{fun } x \rightarrow e_1, e_2) \rangle_Y &= \\
\text{let } \sigma = \langle e_2 \rangle_Y \text{ in } \\
\text{let } \mu = \lambda U. \sum_{s=1}^{N} \sigma(ds)(\langle e \rangle_Y[s/x])^* \text{ in } \\
\text{let } \nu = \lambda U. \mu(U)/\mu(T) \text{ in } \\
(\pi_{\mu}(\nu), \pi_{\sigma}(\nu))
\end{align*}
\]

Figure 16. Probabilistic semantics of \(\mu F\). The semantics is defined only for probabilistic expressions (kindOf(e) = P).

## D Inference
### D.1 Importance Sampling

**Importance sampling.** The most simple inference independently launches \(N\) particles. Each particle executes the importance sampler to compute a pair (result, weight). Results are then normalized in a **categorical distribution**, i.e., a discrete distribution over the results.

The `infer` operator takes a transition function `fun s -> e` and an array of pairs (state, weight) \(S\) of size \(N\) which represents the distribution of possible states across the particles. \(\langle \text{infer}(\text{fun } s \rightarrow e, S) \rangle_Y = \)

\[
\text{let } \mu = \lambda U. \sum_{i=1}^{N} \text{let } s_i, w_i = \langle S \rangle_Y[i] \text{ in } \\
\text{let } s', w' = \langle \text{fun } s \rightarrow e \rangle_Y, w_i(s_i) \text{ in } \\
\overline{w'} * \delta_{\text{dis}(U)}
\]

At each step, the inference executes one step of all the particles and normalizes the scores to return the distribution \(\mu\) of possible results and an updated array of pairs (state, weight) for the next step.
C(let node f x = e) = C(e where
let f_init = \(A(e)\)
let f_step =
  fun (s,x) -> C(e)(s)
C(d1 d2) = C(d1) C(d2)
C(c) = fun s -> (c, s)
C(x) = fun s -> (x, s)
C(last x) = fun s -> (x_last, s)
C(e1, e2) = fun (s1,s2) ->
  let v1, s1' = C(e1)(s1) in
  let v2, s2' = C(e2)(s2) in
  if kindOf e1 then v2, s2' = C(e2)(s2) in
  else let v2, s2' = C(e2)(s2) in
  (v2, (s1', s2'))

C(factor(e)) = fun s ->
  let v, s' = C(e)(s) in
  let _ = factor(v) in (((), (s1', s2')))
Moreover, the compilation of proba nodes introduces an extra argument to the step functions in order to pass the extra information w or (w, g) needed by the inference algorithms.

The code of the inferences algorithms is in the inference directory. The particle filter presented in Section 5.1 is in infer_pf.ml. The entry point of the Delayed Sampling algorithm presented in Section 5.2 is infer_ds_naive.ml and the core of the algorithm is ds_naive_graph.ml. The entry point and the core of the algorithm for the Streaming Delayed Sampling algorithm presented in Section 5.3 are respectively in infer_ds_streaming.ml and ds_streaming_graph.ml.

The Bounded Delayed Sampling algorithm presented in Section 5.2 can be implemented on top of both classical and streaming delayed sampling. The code is in the functor defined in ds_high_level.ml.

Finally, the code for the benchmarks presented Section 6 and Appendix F is available in examples/benchmarks.

Artifact There is an artifact associated to the paper which is available with [1]. It is distributed as a Linux image in the Open Virtualization Format that can be launch using a virtualization player like VirtualBox (https://www.virtualbox.org). The credential to log into the virtual machine are:

user: probzelus
password: probzelus

F Performance Evaluation
This section presents the experimental results. We ran each inference algorithm on a series of benchmarks and measured properties of the execution: accuracy, execution time, memory consumption. All the experiments were run on a server with 32 CPUs (2.60 GHz) and 128 GB memory.

F.1 Benchmarks

Beta-Bernoulli. The Beta-Bernoulli benchmark models an agent that estimates the bias of a coin.

let proba coin (yobs) = xt where
  rec init xt = sample (beta (1., 1.))
  and () = observe (bernoulli xt, yobs)

The model samples xt from a Beta(1,1) distribution, and thereafter evaluates the observations with a Bernoulli distribution of parameter xt. Running SDS on this model is equivalent to exact inference in a Beta-Bernoulli conjugate model [18] where each particle returns the exact solution. The benchmark’s error metric is the mean square error over time between the true coin probability and the expected probability conditioned on the stream of observations.

Gaussian-Gaussian. The Gaussian-Gaussian benchmark models an agent that estimates the mean and the standard deviation of a Gaussian.

let proba gaussian_model (o) = (mu, sigma) where
  rec init mu = sample (gaussian (0., 10.))
  and init sqrt_sigma = sample (gaussian (0., 1.))
  and sigma = sqrt_sigma *. sqrt_sigma
  and () = observe (gaussian (mu, sigma), o)

The initial values for the distribution of the mean mu follows a distribution $\mathcal{N}(0, 10)$ and the distribution of $\sqrt{\text{sigma}}$ is $\mathcal{N}(0, 1)$. The distributions of mu and sigma are conditioned by the observations that follow a distribution $\mathcal{N}(\mu, \sigma)$. In the current implementations of delayed sampling we are doing exact inference only on the mean and not on the standard deviation (even if it would be possible). The benchmark’s error metric is the mean squared error over time between the true mean and standard deviation and the expected probability conditioned on the stream of observations.

Kalman. The Kalman benchmark models an agent that estimates its position based on noisy observations.

let proba delay_kalman (yobs) = xt where
  rec xt = sample (gaussian ((0., 2500.) -> (pre xt, 1.)))
  and () = observe (gaussian (xt, 1.), yobs)

The model chooses an initial position from $\mathcal{N}(0, 2500)$, and chooses subsequent positions from $\mathcal{N}(\text{pre x}, 1)$ where pre x denote the previous position. The model draws the observation at each time step from $\mathcal{N}(x, 1)$ where x is the true position. Running SDS on this model is equivalent to a Kalman filter [25] where each particle returns the exact solution. The benchmark’s error metric is the mean squared error over time between the true position and the expected position conditioned on all previous observations.

Outlier. The Outlier benchmark, adapted from Section 2 of [29], models the same situation as the Kalman benchmark, but with a sensor that occasionally produces invalid readings.

let proba outlier (yobs) = (is_outlier, xt) where
  rec xt = sample (gaussian ((0., 2500.) -> (pre xt, 1.)))
  and init outlier_prob = sample (beta (100., 1000.))
  and is_outlier = sample (bernoulli outlier_prob)
  and () = present is_outlier ->
    observe (gaussian (0., 10000.), yobs)
  else observe (gaussian (xt, 1.), yobs)

The model chooses the probability of an invalid reading from a Beta(100, 10000) distribution, so that invalid readings occur approximately 10% of the time. At each time step, with the previously chosen probability, the model either chooses the observation from the invalid distribution $\mathcal{N}(0, 10000)$, or it chooses the observation from the Kalman model. Running SDS on this model is equivalent to a Rao-Blackwellized particle filter [17] that combines exact inference with approximate particle filtering. The benchmark’s error metric is the mean
The robot starts from the position \( x_0 \) and receives at each step a command \( \text{Right} \) or \( \text{Left} \). It then moves to the left or right following the command with a probability of 10% of remaining in place (modeled by a Bernoulli distribution of parameter 0.1).

The robot maintains a map where each box is a random variable that represents the probability of being black or white (gray level in the Figure 20). The lower line represents the inferred map where the gray level indicates the probability for the cell to be black and the red dots the probability of presence of the robot on the cell.

**Figure 20.** Screenshots of the execution of the SLAM with the PF and SDS inferences. For each screenshot, the top line shows the map, and the blue circle the exact position of the robot. The lower line represents the inferred map where the gray level indicates the probability for the cell to be black or white cell. The robot can move from left to right and observe the color of the cell on which it stands with a sensor. There are two sources of uncertainty: (1) the robot’s wheels are slippery, so the robot can sometimes stay on the spot thinking about moving, (2) the sensor is making read errors, and can reverse the colors. The controller tries to infer the map (color of the cells) and the current position of the robot (Figure 20).

The benchmark’s error metrics is the mean squared error over time between the exact map and position and the expected map and position.

**Multi-Target Tracker.** MTT (Multi-Target Tracker) adapted from [32] is a model where there are a variable number of targets with linear-Gaussian motion models producing linear-Gaussian measurements of the position at each time step. Targets randomly appear according to a Poisson process and each disappear with some fixed probability at each time step. Measurements do not identify which target they came from, and “clutter” measurements that come not from targets but from some underlying distribution add to observations, complicating inference of which measurements are associated to which targets.

We model this with a ProbZelus program that has a state consisting of a list of position-velocity pairs that encode the track of each target. In this example we consider two-dimensional targets, giving us a 4-dimensional vector representing position and velocity together.

The first step is to define helper functions that will be mapped over the list of tracks. The first function tell us how frequently tracks die. They do so with probability \( p_{\text{dead}} \) which we set to \( e^{-0.02} \).

```
let proba death_fn _ = sample (bernoulli (p_dead))
```

We now define how tracks are initialized when they are first created. They are sampled from a multivariate Gaussian distribution with mean \( \mu_{\text{new}} \) set to zero and covariance \( \sigma_{\text{new}} \) set to a diagonal with variance 1 on the positions and variance 0.001 on the velocities.

```
let proba new_track_init_fn _ =
  (new_track_num (),
   sample (mv_gaussian (mu_new, sigma_new)))
```

Next, we define the motion model and update model. Each track \( tr \) is multiplied with the motion matrix \( a_u \) which encodes discrete time integration of position and velocity with time constant 1. We then sample a Gaussian distribution around the new position and velocity with covariance \( \sigma_{\text{update}} \) which is a diagonal matrix with 0.01 variance.
for the position and 0.1 for the variance of the velocity. For the observation model, we project out the position with the projection matrix proj_pos and observe it with covariance matrix sigma_obs which we set to a diagonal of 0.1.

```haskell
let proba state_update_fn (tr_num, tr) =  
    (tr_num,  
     sample (mv_gaussian (a_u *@ tr, sigma_update)))

let observe_fn (_, tr) =  
    (mv_gaussian (proj_pos *@ tr, sigma_obs))

We next define the model for clutter data. We assume that each clutter point is drawn from a multivariate Gaussian with mean mu_clutter which is zero an covariance sigma_clutter which we set to 10.

```haskell
let proba clutter_init_fn _ =  
    (mv_gaussian (mu_clutter, sigma_clutter))
```

The model proceeds as follows. For every track, we use the filter list operator to remove all the tracks that died in this time step. We then sample the number of new tracks n_new from a Poisson distribution with parameter lambda_new which we set to 0.1. After forcing a sample of this value, we use the list constructor ini to build a list of new tracks. We append the survived and new tracks together, and then use the map list operator to apply the motion and observation models to each track. Next, we determine the amount of clutter by subtracting the number of observations from the number of surviving tracks. We then observe that this comes from a Poisson distribution with parameter lambda_clutter which we set to 1.

```haskell
let proba obsfn (var, value) = observe (var, value)
```

```haskell
let proba model inp = t where  
    rec init t = []  
    and t_survived = filter death_fn (last t)  
    and n_new = sample (poisson lambda_new)  
    and t_new = ini new_track_init_fn n_new  
    and t_tot = append t_survived t_new  
    and t = map state_update_fn t_tot  
    and obs = map observe_fn t  
    and n_clutter = (length inp) - (length obs)  
    and () = observe (poisson lambda_clutter, n_clutter)  
    and clutter = ini clutter_init_fn n_clutter  
    and obs_shuffled =  
        sample (shuffle (append obs clutter))  
    and present (not (n_clutter < 0)) ->  
        do () = (iter2 obsfn (obs_shuffled, inp)) done
```

The accuracy metric is based on the Multiple Object Tracking Accuracy [3]. This evaluates whether a track estimate contains the right targets across all time steps within a sufficient tolerance (we set the tolerance to 5 in our example). Conventional MOTA is in [0, 1] with 1 being the best; we have modified it to be in [0, ∞] with 0 being the best by transforming it to MOTA* = 1/MOTA − 1.

Because we estimate a distribution of track estimates, we draw a sample from the track distribution to estimate the expected MOTA*.

**Data.** For each benchmark except Robot and SLAM, we obtained observation data by sampling from the benchmark’s model. In these benchmarks, every run of each benchmark across all experiments uses the same data as input. For SLAM, we pre-sampled the map from the model, but sampled position data on the fly as this data depends on the controller. For the Robot benchmark, we sampled all observations on the fly because they all depend on the command from the controller. This means that for SLAM and Robot, each run uses different position observations.

**F.2 DS vs. PF**

We compare both the accuracy and runtime performance of BDS, SDS, and PF to investigate whether the delayed samplers can achieve better accuracy than the particle filter with the same amount of computational resources.

**Accuracy Methodology.** For a range of selected particle counts, we execute each benchmark multiple times and record the resulting accuracy. To measure accuracy we use the end-to-end error metrics for each benchmark as described in Section F.1. We record the median and the 90% and 10% quantiles after 1000 runs.

**Accuracy Results.** Figures 21 and 23 show the results of the accuracy experiment for the different benchmarks. The error bars show 90% and 10% quantiles, and the center is the median. The vertical lines correspond to the number reported in Figure 10 where there is enough particles to achieve similar accuracy to delayed sampling with 1000 particles. In all cases, SDS is able to achieve equal or better accuracy than BDS which is itself equal or better than PF, but the results vary widely by benchmark. Note that SDS returns the exact posterior distribution for the Coin and Kalman benchmarks therefore its accuracy is independent of the number of particles. On the other-hand, BDS is not exact since the symbolic distributions are sampled at the end of each the step.

**Performance Methodology.** For a range of selected particle counts, we execute each benchmark multiple times (the same number as for the accuracy experiments described above) after a warm-up of 1 run and record the resulting performance: the latency of one step of computation. In the following graphs we report the median latency as well as the 90% and 10% quantiles of the collected data.
**Performance Results.** Figures 22 and 24 shows how the latency for a single step varies with the number of particles for each benchmark. The error bars show 90% and 10% quantiles, and the center is the median. With the three algorithms, the execution time increases linearly with the number of particles. In all cases, PF has lower latency than BDS which has lower latency than SDS.

**Conclusions.** These experiments show that the delayed samplers achieve better accuracy than the particle filter with the same computational resources. For some models SDS is able to compute the exact solution with only one particle (Kalman, Coin). BDS achieves better accuracy when relationships between variables defined in the same step can be exploited (Kalman). At worst the delayed samplers performs as a well as the particle filter (BDS on the Coin, SDS and BDS on the Outlier).

**F.3 SDS vs. DS**

We next evaluate the performance of SDS and BDS relative to our own OCaml implementation of the original delayed sampler (DS). We compare both the performance and memory consumption of the three algorithms at each time step to investigate whether, as the size of the input stream grows large, they can retain constant performance.

**Performance Methodology.** We execute each benchmark 1000 times after a warm-up of 1 run and record the latency. We execute each benchmark with 100 particles (even if only one particle is necessary for DS and SDS on the Coin and Kalman benchmarks to compute the exact distribution) and plot latency as a function of the time step. We report the median latency as well as the 90% and 10% quantiles of the collected data.

**Performance Results.** Figures 25 and 27 shows the latency at each step of a run, aggregated over 1000 runs. PF, BDS, and SDS show nearly constant performance in time but DS gets linearly worse performance for the Kalman and Outlier benchmarks. For the Coin benchmark, the graph of DS remains of constant size because there is only one sample at the first step and then only observe statements.

**Memory Methodology.** We next evaluate the memory consumption of the algorithms. For all benchmarks except the multi-target tracker, memory consumption is deterministic even in the presence of random choices. Therefore, we measure the ideal memory consumption of the execution of each benchmark after each step. The ideal memory consumption is the total amount of live words in the program’s heap. In our implementation, we measure these numbers by forcing a garbage collection after each step. We use OCaml’s standard facilities for forcing garbage collection as well as for measuring the amount of live words. We ran each algorithms 10 times with 100 particles.

For the multi-target tracker, the memory is not deterministic because it is determined by the number of hypothetical tracks, which is random. We report median and 10% and 90% values for memory consumption for this benchmark.

**Memory Results.** Figures 26 and 28 shows the results of the memory consumption experiment. For all benchmarks, PF, BDS, SDS use constant memory over time, including for the multi-target tracker where their memory consumption is random at each time step. However, DS has increasing memory consumption over time for the Kalman, Outlier, and Robot benchmarks. The memory consumption of DS is constant for the Coin benchmark because the graph remains of constant size.

For the multi-target tracker, the memory consumption of DS is based both on the number of hypothesized tracks and the length of the hypothesized tracks. We can see that the memory consumption of DS increases as the first generation of tracks becomes longer, but eventually curtails its memory consumption when these tracks die. MTT’s memory consumption thereafter increases again as the second generation of tracks starts to increase in length.

**Conclusions.** The original DS implementation consumes an increasing amount of memory over time for models that introduce new variables at each step (Kalman, Outlier, and Robot) in contrast to BDS and SDS whose memory consumptions are constant over time. For the multi-target tracker, the DS memory consumption is based on the length of the track which is in principle probabilistically bounded. However, DS still consumes much more memory than PF, BDS, and SDS because the tracks are long-lived.

Furthermore, DS step latency increases without bound as the number of steps becomes large on benchmarks where the memory increases. These observations confirm that the original DS implementation is not practical in a reactive setting.
Figure 21. Accuracy as a function of the number of particles.

Figure 22. Runtime performance as a function of particles.
Figure 23. Accuracy as a function of the number of particles.

Figure 24. Runtime performance as a function of particles.
Figure 25. Runtime performance at each step of a run.

Figure 26. Memory consumption at each step of a run.
Figure 27. Runtime performance at each step of a run.

Figure 28. Memory consumption at each step of a run.