AQUA: Automated Quantized Inference for Probabilistic Programs

Zixin Huang, Saikat Dutta, and Sasa Misailovic

University of Illinois at Urbana-Champaign, IL 61820, USA
{zixinh2,saikatd2,misailo}@illinois.edu

Abstract. We present AQUA, a new probabilistic inference algorithm that operates on probabilistic programs with continuous posterior distributions. AQUA approximates programs via an efficient quantization of the continuous distributions. It represents the distributions of random variables using quantized value intervals (Interval Cube) and corresponding probability densities (Density Cube). AQUA's analysis transforms Interval and Density Cubes to compute the posterior distribution with bounded error. We also present an adaptive algorithm for selecting the size and the granularity of the Interval and Density Cubes.
We evaluate AQUA on 24 programs from the literature. AQUA solved all of 24 benchmarks in less than 43s (median 1.35s) with a high-level of accuracy. We show that AQUA is more accurate than state-of-the-art approximate algorithms (Stan's NUTS and ADVI) and supports programs that are out of reach of exact inference tools, such as PSI and SPPL.

1 Introduction

Many modern applications (e.g., in machine learning, robotics, autonomous driving, medical diagnostics, and financial forecasting) need to make decisions under uncertainty. Probabilistic programming languages (PPLs) offer an intuitive way to model uncertainty by representing complex probabilistic models as simple programs [5]. They expose randomness and Bayesian inference as first-class abstractions by extending standard languages with statements for sampling from probability distributions and probabilistic conditioning. The underlying programming system then automates the intricate details of the probabilistic inference.

Probabilistic inference is a computationally hard problem. Most current approaches that emerged from the statistics and machine learning communities applied aggressive numeric approximations, such as Markov Chain Monte Carlo sampling (MCMC) or Variational Inference (VI). However, these approaches often cannot obtain the level of accuracy that is required in applications such as algorithmic fairness [2], security/privacy [22], sensitivity analysis [13,1], or software testing [8].

Symbolic techniques for inference have been resurging as a more accurate alternative. They use a symbolic representation of the model's state (e.g., elementary functions, piecewise-linear functions, or hypercubes), and compute the posterior distribution algebraically [8,16,19] or closely approximate programs using volume computation [20,2,22]. However, these approaches are often limited by the classes of programs they can solve. For instance, continuous programs pose a major challenge for

these approaches due to integrals in posterior calculation. State-of-the-art symbolic solvers cannot solve many integrals exactly (often, the integrals do not have a closed form). Similarly, volume computation approaches have a limited support for continuous distributions (e.g., do not allow for conditioning on continuous random variables) and/or compute the probability of a single event, not the entire posterior distribution. *An intriguing research question is how to approximate multi-dimensional continuous distributions in a principled manner that allows for more expressive programs and can solve programs that are out of reach of existing tools for exact inference.*

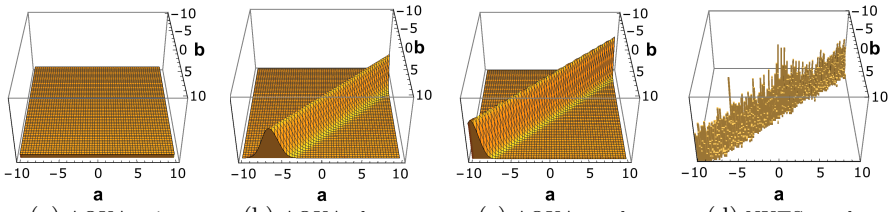
AQUA. We present AQUA, a novel system for symbolic inference that uses quantization of probability density function for delivering scalable and precise solutions for a broad range of probabilistic programs. AQUA's inference algorithm approximates the original continuous program via an efficient quantization of the continuous distributions by using multi-dimensional tensor representations that we call *Interval Cube and Density Cube*. The Interval Cube stores the quantized value ranges of variables in the probabilistic program. The Density Cube approximates the joint posterior distribution by recording the probability of each hypercube contained in the interval cube.

AQUA's analysis transforms the symbolic state to compute quantized approximate posterior distribution.We derive the bounds for the approximation error (due to the quantization and integration) and show that our inference converges in distribution to the true posterior. We also present an adaptive algorithm for automatically selecting the granularity of the Interval and Density Cubes.

Example. Figure 1 presents a probabilistic program that represents the distribution of two random variables. In the program, we have two random variables `a` and `b`, each having `uniform` *prior* distribution (Lines 3-4). We then condition the model on 40 data points `Y`, assuming that each point is normally distributed with the mean `a+b` (Lines 5-6). We finally query for the joint *posterior* distribution (i.e., the distribution of latent variables `a` and `b` after observing the data on Line 6).

```
1 D=40
2 Y=[3.4,0.3,...]
3 a~uniform(-10,10)
4 b~uniform(-10,10)
5 for (i in 1:D)
6  Y[i]~normal(a+b,1)
7 return a,b
```

Fig. 1: Example



(a) AQUA prior (b) AQUA observe (c) AQUA result (d) NUTS result

Fig. 2: AQUA estimated probabilistic density and NUTS histogram

Figure 2 presents AQUA's results: (a) shows the prior of the two variables, (b) shows the likelihood (observation) on a single data point, and (c) shows the posterior distribution. On each plot, the X-axis and Y-axis represent `a` and `b` values, and the Z-axis values are the probability densities computed by AQUA. AQUA computes the result in 0.76s, whereas an MCMC based inference algorithm (NUTS) produces a less accurate posterior within the same amount of time (Figure 2(d)).

Evaluation. We evaluate our implementation of AQUA on a set of 24 probabilistic programs from the literature. We compare AQUA with exact inference – PSI [8] and SPPL [19] – and approximate inference – MCMC and VI implemetations in Stan [5]. We show that AQUA can solve programs that are out of reach for PSI and SPPL. Our results show AQUA solved all benchmarks in less than 43s (median 1.35s). It is significantly more accurate than VI for all programs (for the Kolmogorov-Smirnov metric). AQUA is substantially more accurate than MCMC for 10 programs, even when MCMC is given substantially more time to complete. We also present a case study that shows AQUA can precisely capture the tails of the distribution of robust models.

Contributions. This paper makes the following contributions:

- **Inference Algorithm:** We present AQUA, a novel inference algorithm that works on general, real-world probabilistic programs with continuous distributions based on quantization and symbolic computation.
- **Quantization with Interval and Density Cubes:** Our analysis defines symbolic transformers on the abstract state consisting of the Interval and Density Cubes. We also present theoretical bounds on the quality of approximation.
- **Inference Algorithm Optimizations:** We present algorithm extensions that automatically refine the size/granularity of the analysis to satisfy a given precision threshold and aggressively reduce the analysis overhead of local variables.
- **Evaluation:** Our experiments show that AQUA is more accurate than approximate inference algorithms (Stan's MCMC/VI) and supports programs with conditioning on continuous distributions that are out of reach of exact inference tools (PSI and SPPL).

2 Preliminaries

Language Syntax and Semantics. Figure 3 describes the syntax of a probabilistic program using an imperative, first-order intermediate representation, drawing from Storm-IR [6,7]. It has statements for sampling from distributions[1] and conditioning on data with `factor` and `observe`.

$x \in$ Vars
$c \in$ Consts
$op \in \{+,-,*,>,\ldots\}$
$d \in$ {`Normal`,`Uniform`,...}

`E` := c | x | `E[E*]` | `E` op `E` | d`(E*).pdf(E*)` | f`(E*)`
`S` := x `=` `E` | $x \sim$ d`(E`*`)` | `factor(E)` | `observe(d(E`*`),`x`)`
| `if (E) S`* `else S`* | `for` $x \in$ `1..N; {S`*`}`
`P` := `S`$^+$`; return` x^+

Fig. 3: Syntax of AQUA's language

The language semantics are standard, inspired by those presented by Gorinova et al. [11] (We present the detailed semantics rules in the Appendix, available in the full version of the paper). In summary, a probabilistic program evaluates the *posterior probability density function*. Our operational semantics for a program

[1] We support common continuous distributions including Normal, Uniform, Exponential, Beta, Gamma, Student-T, Laplace, Triangular, and any mixture of the above distributions.

defines its effect on the program state, σ, which maps variables to values. A value V can either be a constant c or an array of values $[c_1,c_2,...]$. The notations $\sigma(x)$ and $\sigma(x \mapsto V)$ denote accessing and updating a variable x respectively. We refer to the return variables of the program as the *global variables*, and the others as *local variables*. We allow local variables to have discrete distributions (e.g. Bernoulli), as long as the density of the global variables are Lipschitz continuous. We define a special variable $\mathcal{L} \in \mathbb{R}^+$ which tracks the *unnormalized posterior density* of the probabilistic program. We initialize $\sigma(\mathcal{L})$ to 1.0 at the start of the program.

Probability Density. We review several basic terms from the probability theory. Let $\boldsymbol{x}$ be the set of variables with values in V, and $\mathcal{D}$ be the set of observed data points. Then, the posterior probability density function is $p(\boldsymbol{x}|\mathcal{D}) : V \to \mathbb{R}$, such that $\int_{\boldsymbol{x} \in V} p(\boldsymbol{x}|\mathcal{D}) d\boldsymbol{x} = 1$. The distribution $p(\boldsymbol{x}|\mathcal{D})$ can be calculated from the *unnormalized probability density function* $f(\boldsymbol{x},\mathcal{D}) : V \to \mathbb{R}$, by $p(\boldsymbol{x}|\mathcal{D}) = \frac{1}{z} f(\boldsymbol{x},\mathcal{D})$, where z is the normalizing constant: $z = \int f(\boldsymbol{x},\mathcal{D}) d\boldsymbol{x}$. If $\boldsymbol{x}_{-i}$ contains all the variables in $\boldsymbol{x}$ excluding x_i, we define the *marginal probability density function* of x_i as $p(x_i|\mathcal{D}) = \int p(\boldsymbol{x}|\mathcal{D})\ d\boldsymbol{x}_{-i}$. Hereon, we omit data symbol $\mathcal{D}$ to write $p(\boldsymbol{x})$ and $f(\boldsymbol{x})$ when clear from the context. In the semantics, $f(\boldsymbol{x})$ is represented by $\sigma(\mathcal{L})$.

3 AQUA's Probabilistic Inference using Density Cubes

3.1 Notations and Basic Definitions

We represent the closed, bounded set $\{x \in \mathbb{R} | a \leq x \leq b\}$ with its lower-bound $a \in \mathbb{R}$ and upper-bound $b \in \mathbb{R}$. We denote this abstraction as an **interval** $I = [a,b] \in \mathbb{R}^2$. We refer to the lower and upper bound of I as $\underline{I}$ and $\overline{I}$, respectively ($\underline{I},\overline{I} \in \mathbb{R}$).

A probabilistic program lifts a normal program operating on single values to a *distribution* over values. Hence, a probabilistic program represents a joint distribution over its variables. For our symbolic analysis, to represent the quantized values of variables, we define tensors of intervals which we will refer to as *Interval Cube*. We also assign a probability density to each interval in the Interval Cube. We will refer to this assignment of densities as *Density Cube*. If there are N variables in the program, the Density Cube will be an N-dimensional tensor.

Definition 1 (Interval Cube)**.** We represent the value of a variable x with Interval Cube, $\boldsymbol{I}^x_{M_1,M_2,...,M_N}$ where $[M_1,M_2,...,M_N]$ represents the shape of the Interval Cube and each $M_i \in \mathbb{N}$ is the *number of intervals* (*splits*) along the i-th dimension. Each element of $\boldsymbol{I}^x_{M_1,M_2,...,M_N}$ is a single interval. We let $\mathbb{I}$ be the set of all Interval Cubes. For a constant c, we denote its Interval Cube as $[c]$, meaning a singleton interval with both lower and upper bounds being c.

To simplify the notation, we hereon denote the shape of the hypercube as $\boldsymbol{M} = [M_1,M_2,...,M_N]$ and each index in the hypercube is $\boldsymbol{m} \in \mathbb{M}$, $\mathbb{M} = \{[m_1,...,m_N] | m_i \in [1,...,M_i],\ i \in \{1,...,N\}\}$. We write $\boldsymbol{K} = \boldsymbol{M}_1 \odot \boldsymbol{M}_2$ as the element-wise product (Hadamard product) of two shape vectors, namely $K_i = M_{1i} \times M_{2i}$, $i \in \{1,...,N\}$. We use $\boldsymbol{m}_1$ to denote the index of a Interval Cube with shape $\boldsymbol{M}_1$, $\boldsymbol{m}_1 = [m_1,...,m_N]$, $m_i \in \{1,...,M_{1i}\}$, and similarly we use $\boldsymbol{m}_2$ for index in $\boldsymbol{M}_2$, and $\boldsymbol{k}$ for index in $\boldsymbol{K}$.

Table 1: Correspondence of Symbolic Analysis and Concrete Analysis

Concrete	Symbolic
Value $\sigma(\boldsymbol{x})$	Interval Cube $\sigma^{\#}(\boldsymbol{x})$
Density $\sigma(\mathcal{L})$	Density Cube $\sigma^{\#}(P)$
State : $(\mathcal{P}(\textbf{Vars} \mapsto \textbf{Value}) \mapsto [0,1])$	**Astate** : $(\mathcal{P}(\textbf{Vars} \mapsto$ Interval Cube$) \mapsto$ Density Cube$)$
$[\![E]\!]$: **State**$\mapsto$**Value**	$[\![E]\!]^{\#}$: **Astate**$\mapsto$Interval Cube
$[\![S]\!]$: **State**$\mapsto$**State**	$[\![S]\!]^{\#}$: **Astate**$\mapsto$**Astate**

Definition 2 (Density Cube)**.** For a given probabilistic program *Prog* with N variables, we define the Density Cube with shape $\boldsymbol{M}$ as $\boldsymbol{P}_{\boldsymbol{M}}^{Prog}$, where

$$\boldsymbol{P}_{\boldsymbol{M}}^{Prog}(\boldsymbol{m}) = p_{\boldsymbol{m}}, \text{ for each index } \boldsymbol{m} \in \mathbb{M},$$

and $p_{\boldsymbol{m}}$ denotes the value of the unnormalized probability density function at the lower bound of the corresponding interval in the Interval Cube. The densities at the lower bound of intervals will help us do numerical integration for posterior calculation. Further, $\boldsymbol{P}_{\boldsymbol{M}}^{Prog} \in \mathbb{R}^{\boldsymbol{M}}$, and $p_{\boldsymbol{m}} \in \mathbb{R}$.

Definition 3 (Symbolic Domain)**.** Our *symbolic state* has two components, a map from variables to Interval Cubes, and a Density Cube representing the joint density approximation. Let **Var** denote the set of variables, and $\mathcal{P}$ be the power set, the domain of the symbolic state is $\Sigma = \mathcal{P}(\text{Var} \mapsto \mathbb{I}) \times \mathbb{R}^{\boldsymbol{M}}$ a symbolic state $\sigma^{\#} \in \Sigma$ will have the form

$$\sigma^{\#} = \left\langle \{x_1 \mapsto \boldsymbol{I}_{\boldsymbol{M}_1}^{x_1}, x_2 \mapsto \boldsymbol{I}_{\boldsymbol{M}_2}^{x_2}, ..., x_i \mapsto \boldsymbol{I}_{\boldsymbol{M}_i}^{x_i}, ...\}, P \mapsto \boldsymbol{P}_{\boldsymbol{M}}^{Prog} \right\rangle.$$

The symbolic domain expresses a piecewise constant interpolation of the joint probability density at a program point. Hereon, we refer to the set of all the variables in the state $\sigma^{\#}$ as $\boldsymbol{x} = \{x_1, x_2, ..., x_N\}$.

3.2 Analysis

We approximate the posterior density function of variables in our symbolic states. Table 1 presents the correspondence of the objects in concrete semantics to symbolic states. While a concrete state has a single valuation of variables and evaluates to a single density value, our symbolic state stores all possible variable values in Interval Cube and corresponding probability densities in Density Cube. As the concrete semantics for a expression maps state to values, the symbolic semantics map symbolic state to Interval Cube; and as the concrete semantics for a statement map state to state, our symbolic semantics map symbolic state to symbolic state.

Analysis of Expressions. The symbolic transformer $[\![E]\!]^{\#}$ on an expression E takes a symbolic state $\sigma^{\#}$: **Astate** as input, and outputs an Interval Cube. Figure 4 presents the rules. We explain two important cases in detail:

• $[\![E_1 \ op \ E_2]\!]^{\#}$: For the arithmetic/boolean operation on two Interval Cubes, which may not always have the same shape, the resulting Interval Cube needs to contain all possible value combinations. Specifically, for $\boldsymbol{I}_{\boldsymbol{M}_1}^{E_1}$ with shape $\boldsymbol{M}_1 = [M_{11}, ..., M_{1N}]$ and $\boldsymbol{I}_{\boldsymbol{M}_2}^{E_2}$ with shape $\boldsymbol{M}_2 = [M_{21}, ..., M_{2N}]$, the result $\boldsymbol{I}_{\boldsymbol{K}}^{E_1 \ op \ E_2}$ has shape $\boldsymbol{K} = [K_1, ..., K_N]$ with $K_i = M_{1i} \times M_{2i}$ to capture all the combinations of elements from

$$
\begin{aligned}
[\![E]\!]^{\#} &\mapsto (\textbf{Astate} \mapsto \text{Interval Cube}) \\
[\![x]\!]^{\#} &\coloneqq \lambda\sigma^{\#}.\sigma^{\#}(x) \\
[\![c]\!]^{\#} &\coloneqq \lambda\sigma^{\#}.[c] \\
[\![E_1[E_2]]\!]^{\#} &\coloneqq \lambda\sigma^{\#}.\text{let } [c,c] = [\![E_2]\!]^{\#}\sigma^{\#} \text{ in } [\![E_1[c]]\!]^{\#}\sigma^{\#} \\
[\![E_1 \ op \ E_2]\!]^{\#} &\coloneqq \lambda\sigma^{\#}.\text{let } \boldsymbol{I}^{E_1}_{\boldsymbol{M}_1} = [\![E_1]\!]^{\#}\sigma^{\#},\ \boldsymbol{I}^{E_2}_{\boldsymbol{M}_2} = [\![E_2]\!]^{\#}\sigma^{\#},\ \boldsymbol{K} = \boldsymbol{M}_1 \odot \boldsymbol{M}_2 \\
&\qquad \text{in } \boldsymbol{I}^{E_1 \ op \ E_2}_{\boldsymbol{K}}, \text{ where } \boldsymbol{I}^{E_1 \ op \ E_2}_{\boldsymbol{K}}(\boldsymbol{k}) = \boldsymbol{I}^{E_1}_{\boldsymbol{M}_1}(\boldsymbol{m}_1) \ op \ \boldsymbol{I}^{E_2}_{\boldsymbol{M}_1}(\boldsymbol{m}_2) \\
[\![d(E_1,...,E_{n-1}).\text{pdf}(E_n)]\!]^{\#} &\coloneqq \lambda\sigma^{\#}.\text{let } \boldsymbol{I}^{E_1}_{\boldsymbol{M}_1} = [\![E_1]\!]^{\#}\sigma^{\#},...,\boldsymbol{I}^{E_n}_{\boldsymbol{M}_n} = [\![E_n]\!]^{\#}\sigma^{\#},\ \boldsymbol{K} = \bigodot_{i=1}^{n} \boldsymbol{M}_i, \\
&\qquad \text{in } \boldsymbol{I}^{\text{dpdf}}_{\boldsymbol{K}}, \text{ where } \boldsymbol{I}^{\text{dpdf}}_{\boldsymbol{K}}(\boldsymbol{k}) = d\text{\_pdf}(\boldsymbol{I}^{E_n}_{\boldsymbol{M}_n}(\boldsymbol{m}_n), \boldsymbol{I}^{E_1}_{\boldsymbol{M}_1}(\boldsymbol{m}_1),...,\boldsymbol{I}^{E_{n-1}}_{\boldsymbol{M}_{n-1}}(\boldsymbol{m}_{n-1}))
\end{aligned}
$$

Fig. 4: Analysis of Expressions

$\boldsymbol{I}^{E_1}_{\boldsymbol{M}_1}$ and $\boldsymbol{I}^{E_2}_{\boldsymbol{M}_2}$. If $\boldsymbol{M}_1$ and $\boldsymbol{M}_2$ are not of the same length, we reshape both $\boldsymbol{I}^{E_1}_{\boldsymbol{M}_1}$ and $\boldsymbol{I}^{E_2}_{\boldsymbol{M}_2}$ to have the same dimension, by letting some M_{1i} or M_{2i} to have value 1. We let the arithmetic or boolean operation on the interval pairs be $\boldsymbol{I}^{E_1}_{\boldsymbol{M}_1}(\boldsymbol{m}_1)$ *op* $\boldsymbol{I}^{E_2}_{\boldsymbol{M}_2}(\boldsymbol{m}_2) \coloneqq [\underline{\boldsymbol{I}^{E_1}_{\boldsymbol{M}_1}(\boldsymbol{m}_1)} \ op \ \underline{\boldsymbol{I}^{E_2}_{\boldsymbol{M}_2}(\boldsymbol{m}_2)}, \overline{\boldsymbol{I}^{E_1}_{\boldsymbol{M}_1}(\boldsymbol{m}_1)} \ op \ \overline{\boldsymbol{I}^{E_2}_{\boldsymbol{M}_2}(\boldsymbol{m}_2)}]$. We handle the case with multiple intervals analogously. This operation on multiple Interval Cubes can be implemented efficiently with the *broadcast* function in tensor libraries.

• $[\![d(E_1,...,E_{n-1}).\text{pdf}(E_n)]\!]$: Similar to arithmetic operator, we apply the mathematical density d_pdf(_) of the distribution d whose parameters (e.g., mean, location, shape or variance) are intervals obtained by evaluating $E_1,...,E_{n-1}$, and it takes the intervals of E_n for which we seek the density. We denote the shape of the result Interval Cube as $\boldsymbol{K}$, which is computed from the shape of the input Interval Cubes.

Analysis of Statements. Figure 5 presents the transformers $[\![S]\!]^{\#}$ on statements S, which takes an abstract state $\sigma^{\#}$: **Astate** as input, and outputs an abstract state. We explain two important rules where we modify Density Cube (the remaining statements are standard or rely on these two rules):

• $[\![x \sim d(E_1,...,E_n)]\!]^{\#}$, $[\![\text{factor}(E)]\!]^{\#}$: We first evaluate d.pdf(_) of E into Interval Cube, and multiply the current Density Cube with the lower bound of intervals from the Interval Cube. Then at the lower bound of each interval, the density is the same as the one from concrete semantics (Lemma 7). Intuitively, we discretize the density function and use the density at the lower bound to represent each interval. For convenience, our discretization uses the density at the lower bound. Using the density at the upper bound or the midpoint is also possible, and our accuracy guarantee (Theorem 10) still holds.

• $[\![\text{if } (E) \text{ then } \{S_1\} \text{ else } \{S_2\}]\!]^{\#}$: We first solve the result from two branches one conditioning on E and the other on $1-E$. The true boolean expressions evaluate to 1 and false to 0 in our analysis, and we get the interval cubes for E and $1-E$ from expression rules (Figure 4). We then *Join* the result states by adding up the Density Cubes from both branches.

Definition 4 (Joins)**.** *Join* ($\sqcup$) adds the Density Cubes from two states. Formally, $\sigma^{\#}_1 \sqcup \sigma^{\#}_2 = \sigma^{\#}_1(P \mapsto \boldsymbol{P}^{Prog}_{M_1,M_2,...,M_N})$, where each element in $\boldsymbol{P}^{Prog}_{M_1,M_2,...,M_N}$ at

$$
\begin{aligned}
[\![S]\!]^{\#} &\mapsto (\mathbf{Astate} \mapsto \mathbf{Astate}) \\
[\![\text{skip}]\!]^{\#} &\coloneqq \lambda\sigma^{\#}.\sigma^{\#} \\
[\![S_1;S_2]\!]^{\#} &\coloneqq \lambda\sigma^{\#}.[\![S_2]\!]^{\#}([\![S_1]\!]^{\#}\sigma^{\#}) \\
[\![x = E]\!]^{\#} &\coloneqq \lambda\sigma^{\#}.\text{let } \boldsymbol{I} = [\![E]\!]^{\#}\sigma^{\#} \text{ in } \sigma^{\#}(x \mapsto \boldsymbol{I}) \\
[\![x \sim d(E_1,...,E_n)]\!]^{\#} &\coloneqq \lambda\sigma^{\#}.\text{let } \boldsymbol{P}_{\boldsymbol{M}_0} = \sigma^{\#}(P),\ \boldsymbol{I}_{\boldsymbol{K}}^{\text{dpdf}} = [\![d(E_1,...,E_n).\text{pdf(x)}]\!]^{\#}\sigma^{\#}, \text{ in} \\
&\qquad \text{let } \boldsymbol{M} = \boldsymbol{M}_0 \odot \boldsymbol{K}, \text{ in } \sigma^{\#}(P \mapsto \boldsymbol{P}'_{\boldsymbol{M}}), \\
&\qquad \text{where } \boldsymbol{P}'_{\boldsymbol{M}}(\boldsymbol{m}) = \boldsymbol{P}_{\boldsymbol{M}_0}(\boldsymbol{m}_0) \cdot \underline{\boldsymbol{I}_{\boldsymbol{K}}^{\text{dpdf}}(\boldsymbol{k})}, \text{ for all } \boldsymbol{m} = \boldsymbol{m}_0 \odot \boldsymbol{k}, \\
&\qquad \boldsymbol{m}_0 \in \{[m_{01},...,m_{0N}] | m_{0i} \in \{1,...,M_{0N}\}\}, [M_{01},...,M_{0N}] = \boldsymbol{M}_0, \\
&\qquad \boldsymbol{k} \in \{[k_1,...,k_N] |\ k_i \in \{1,...,K_N\}\},\ [K_1,...,K_N] = \boldsymbol{K} \\
[\![\text{factor}(E)]\!]^{\#} &\coloneqq \lambda\sigma^{\#}.\text{let} \boldsymbol{P}_{\boldsymbol{M}_0} = \sigma^{\#}(P),\ \boldsymbol{I}_{\boldsymbol{K}} = [\![E]\!]^{\#}\sigma^{\#},\ \boldsymbol{M} = \boldsymbol{M}_0 \odot \boldsymbol{K} \\
&\qquad \text{in } \sigma^{\#}(P \mapsto \boldsymbol{P}'_{\boldsymbol{M}}), \text{ where } \boldsymbol{P}'_{\boldsymbol{M}}(\boldsymbol{m}) = \boldsymbol{P}_{\boldsymbol{M}_0}(\boldsymbol{m}_0) \cdot \underline{\boldsymbol{I}_{\boldsymbol{K}}(\boldsymbol{k})} \\
&\qquad \text{where } \boldsymbol{P}'_{\boldsymbol{M}}, \boldsymbol{P}_{\boldsymbol{M}_0} \text{and } \boldsymbol{P}_{\boldsymbol{K}} \text{ are as above} \\
[\![\text{observe}(d(E_1,...,E_n),x)]\!]^{\#} &\coloneqq \lambda\sigma^{\#}.[\![\text{factor}(d(E_1,...,E_n).\text{pdf}(x))]\!]^{\#}\sigma^{\#} \\
[\![\text{if } (E) \text{ then } \{S_1\} \text{ else } \{S_2\}]\!]^{\#} &\coloneqq \lambda\sigma^{\#}.\left([\![\text{factor}(E);S_1]\!]^{\#}\sigma^{\#}\right) \sqcup \left([\![\text{factor}(1\text{-}E);S_2]\!]^{\#}\sigma^{\#}\right) \\
[\![\text{for } (i \text{ in } E_1..E_2)\ S]\!]^{\#} &\coloneqq \lambda\sigma^{\#}.[\![i = E_1; \text{if } (i \leq E_2) \text{then} \{S; \text{for } (i \text{ in } E_1 + 1..E_2) S\} \text{else} \{\text{skip}\}]\!]^{\#}\sigma^{\#}
\end{aligned}
$$

Fig. 5: Analysis of Statements

location $\boldsymbol{m}$ is $\sigma_1^{\#}(P)(\boldsymbol{m}) + \sigma_1^{\#}(P)(\boldsymbol{m})$, with $\boldsymbol{m} = [m_1, m_2, ..., m_N]$, $m_i \in \{1,...,M_i\}$. Since we already initialized the global variables with their Interval Cube, $\sigma_1^{\#}$ and $\sigma_2^{\#}$ should have the same variables and Interval Cubes. Then the joint probability density P is changed to the sum of the densities from both states. Similarly, we can define *Meet* ($\sqcap$) by product of $\sigma_1^{\#}(P)(\boldsymbol{m})$ and $\sigma_1^{\#}(P)(\boldsymbol{m})$.

Algorithm. Algorithm 1 takes as input a probabilistic program *Prog*, the shape vector $\boldsymbol{M}$ where each element M_i is the number of intervals for variable x_i, and the interval bounds $\boldsymbol{C}$ (optional). In Section 4, we describe an adaptive scheme to automatically search for a proper $\boldsymbol{C}$ for the analysis.

First, it initializes the probability density variable P with the single interval [1.0] (Line 2). Then, it splits the value domain for each x_i in *SampledVars*, which are variables sampled from a prior distribution $x_i \sim d(E_1,...,E_n)$ and not from deterministic assignments, into M_i equi-length intervals in C_i (in the function *GetInitIntervals*, Line 3-5). M_i is the i-th element in $\boldsymbol{M}$, and C_i is the i-th element in $\boldsymbol{C}$.

The algorithm follows the analysis rules to get the state at the end of the program (Line 6). Then it computes the joint probability density estimation $\hat{f}$, as a piecewise function of $\sigma^{\#}(P)$ (Line 7).

The result $\hat{f}(\boldsymbol{x})$ is an approximation of the true unnormalized probability density function $f(\boldsymbol{x})$. In the concrete domain, the posterior probabilistic density function is calculated as $p(\boldsymbol{x}) = \frac{1}{z} f(\boldsymbol{x})$, but the integration $z = \int f(\boldsymbol{x}) d\boldsymbol{x}$ is often intractable. We compute our approximation $\hat{z}$ using integration on the piecewise function:

Definition 5 (Integration for Normalizing Constant)**.** Suppose there are N sampled variables $\boldsymbol{x}$ in the program, and let $\boldsymbol{C} = \bigotimes_{i=1}^{N} [a_i, b_i] \subset \mathbb{R}^N$ for each $x_i \in [a_i, b_i] \subset \mathbb{R}$ be

Algorithm 1 Posterior Interval Analysis Algorithm

1: **procedure** PosteriorAnalysis(*Prog*, $\boldsymbol{M}$, $\boldsymbol{C}$)
2: $\sigma^{\#}_{init} \leftarrow \{P \mapsto [1]\}$ ▷ Initialize with probability 1
3: **for** $x_i \in SampledVars(Prog)$ **do**
4: $\sigma^{\#}_{init}[x_i] \leftarrow GetInitIntervals(x_i, M_i, C_i)$
5: **end for**
6: $\sigma^{\#} \leftarrow [\![Prog]\!]\sigma^{\#}_{init}$ ▷ Apply analysis rules
7: $\hat{f}(\boldsymbol{x}) \leftarrow PiecewiseFunc(\sigma^{\#}(P))$
8: $\hat{z} \leftarrow \int_{\sigma^{\#}[\boldsymbol{x}]} \hat{f}(\boldsymbol{x})\, d\boldsymbol{x}$; $\hat{p}(\boldsymbol{x}) \leftarrow \frac{1}{\hat{z}} \hat{f}(\boldsymbol{x})$ ▷ Normalize the Posterior
9: **for** $x_i \in SampledVars(Prog)$ **do**
10: $Marginal[x_i] \leftarrow \frac{1}{z} \int_{\sigma^{\#}[\boldsymbol{x}_{-i}]} \hat{p}(\boldsymbol{x})\, d\boldsymbol{x}_{-i}$ ▷ Marginalize
11: **end for**
12: **return** ($\hat{p}$, *Marginal*)
13: **end procedure**

the bounded domain used in the analysis ($\bigotimes$ represents the Cartesian Product on intervals on $\mathbb{R}$). We initialize $\sigma^{\#}[\boldsymbol{x}] = \boldsymbol{C}$ in the analysis. Then $z = \int_{\boldsymbol{C}} f(\boldsymbol{x})\, d\boldsymbol{x}$ is approximated with $\hat{z} = \int_{\sigma^{\#}[\boldsymbol{x}]} \hat{f}(\boldsymbol{x})\, d\boldsymbol{x} = \sum_{\boldsymbol{m} \in \mathbb{M}} (\prod_{i=1}^{N} (\overline{\boldsymbol{I}^{x_i}_{M_i}(\boldsymbol{m})} - \underline{\boldsymbol{I}^{x_i}_{M_i}(\boldsymbol{m})}) \cdot \boldsymbol{P}^{P}_{\boldsymbol{M}}(\boldsymbol{m}))$.

The algorithm finally computes the posterior and the marginals for every variable (Lines 8-11). When the program has N variables, and each variable has the same number of intervals M, Algorithm 1 has the time complexity $\mathcal{O}(N \cdot M^N)$ and space complexity $\mathcal{O}(M^N)$.

3.3 Formal Guarantee of Accuracy

In this section we formally derive how well the symbolic state $\sigma^{\#}$ approximates the joint unnormalized density function f and the posterior density function p.

Definition 6 (Concretization of Symbolic States)**.** Define γ as the concretization function, s.t. $\gamma(\sigma^{\#}) = \hat{f}$, where $\hat{f}(\boldsymbol{x}) = \sigma^{\#}(P)(\boldsymbol{m})$ if $\boldsymbol{x} \in \bigotimes_{i=1}^{N} [\underline{\boldsymbol{I}^{x_i}_{M_i}(\boldsymbol{m})}, \overline{\boldsymbol{I}^{x_i}_{M_i}(\boldsymbol{m})}) \subset \mathbb{R}^N$ for any $\boldsymbol{m}$, and 0 otherwise.

Lemma 7 shows that at any program point, the error is bounded if we use the analysis result $\gamma(\sigma^{\#}) = \hat{f}$ as an approximation of joint density function f, and the error will reduce by the more number of intervals. To simplify the presentation, we use $\underline{\boldsymbol{x}}^{(\boldsymbol{m})} = [\underline{\boldsymbol{I}^{x_1}_{M_1}(\boldsymbol{m})}, ..., \underline{\boldsymbol{I}^{x_N}_{M_N}(\boldsymbol{m})}]$ for all variables, and analogously for $\overline{\boldsymbol{x}}^{(\boldsymbol{m})}$.

Lemma 7 (Discretization Error)**.** *The error of discretization is* $|\hat{f}(\boldsymbol{x}) - f(\boldsymbol{x})| \leq \mu \cdot \max_{\boldsymbol{m}} \|\overline{\boldsymbol{x}}^{(\boldsymbol{m})} - \underline{\boldsymbol{x}}^{(\boldsymbol{m})}\|$ *if* $\boldsymbol{x} \neq \underline{\boldsymbol{x}}^{(\boldsymbol{m})}$*, and if* $\boldsymbol{x} = \underline{\boldsymbol{x}}^{(\boldsymbol{m})}$ *the error is 0.*

Proof Sketch. We show that at any program point, (1) $\sigma^{\#}(P)(\boldsymbol{m}) = f(\boldsymbol{x})$ when $\boldsymbol{x} = \underline{\boldsymbol{x}}^{(\boldsymbol{m})}$, meaning the abstract transformers are exact at the lower bounds, and (2) f is μ-Lipschitz continuous. By definition of μ-Lipschitz continuous, $|f(\boldsymbol{x}_1) - f(\boldsymbol{x}_2)| \leq$

$\mu \cdot \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|$, we can prove the Lemma. The proof is by structural induction: we first show at initialization of the program, $\sigma^{\#}$ satisfies (1) and (2) because $f(\boldsymbol{x}) = 1.0$ and $\sigma^{\#}(P)(\boldsymbol{m}) = [1.0]$. Then for each statement, we show if the pre-state satisfies (1), the post-state has $\sigma^{\#}(P)(\boldsymbol{m}) = f(\underline{\boldsymbol{x}}^{(\boldsymbol{m})})$; and if the pre-state satisfies (2), f is Lipschitz continuous. We present the full proof in the Appendix. □

The error of AQUA's approximation to the normalizing constant z is also bounded:

Lemma 8 (Integration Error)**.** *Let $U = \prod_{i=1}^{N}(b_i - a_i)$ be the volume of $\boldsymbol{C}$. For all the probability distributions supported in our language, the error is $|z - \hat{z}| \leq U\mu \max_{\boldsymbol{m}} \|\overline{\boldsymbol{x}}^{(\boldsymbol{m})} - \underline{\boldsymbol{x}}^{(\boldsymbol{m})}\|$. If we use M equal-length intervals for each variable, $|z - \hat{z}| \leq U\mu \frac{1}{M}(\sum_{i=1}^{N}(b_i - a_i)^2)^{\frac{1}{2}}$. Then $|z - \hat{z}| \to 0$ as $M \to \infty$.*

Proof Sketch. Recall, all posteriors f in our language (Section 2) are Lipschitz continuous. We derive the error bound by applying the Lipschitz continuous property of f and the triangle inequality. We present the full proof in the Appendix. □

Moreover, the integration error bound above will decrease when we decrease the interval length, or increase the number of intervals. Then at the end of the analysis, we approximate the *posterior probability density function* $p(\boldsymbol{x})$ on $\boldsymbol{C}$ as:

Definition 9 (Posterior Probability Density Approximation)**.** Define $\hat{p}(\boldsymbol{x}) = \frac{1}{\hat{z}}\hat{f}(\boldsymbol{x})$ as the approximation of $p(\boldsymbol{x})$.

Now we show the end-to-end error of the analysis. As Theorem 10 states, by applying sufficiently many intervals, the random variables following AQUA's posterior estimation in $\boldsymbol{C}$ will *converge in distribution* to the true posterior in C. Without loss of generality, suppose we apply at least M equal-length intervals for each variable in its domain $[a_i, b_i]$, i.e. $M = \min\{M_1, M_2, ..., M_N\}$. And we refer $\hat{p}_M(\boldsymbol{x})$ as AQUA's approximation of $p(\boldsymbol{x})$ by applying at least M equal-length intervals for each variable.

Theorem 10 (Convergence of Posterior Density Approximation)**.** *Define $F_{\boldsymbol{C}}(\boldsymbol{x}) = \frac{1}{z}\int_{-\boldsymbol{\infty}}^{\boldsymbol{x}} \mathbf{1}_{[\boldsymbol{u} \in \boldsymbol{C}]} \cdot p(\boldsymbol{u}) d\boldsymbol{u}$ as the true cumulative distribution function (CDF) on $\boldsymbol{C}$, where $z = \int_{\boldsymbol{C}} p(\boldsymbol{x}) d\boldsymbol{x}$, and $\hat{F}_{\boldsymbol{C},M}(\boldsymbol{x}) = \int_{-\boldsymbol{\infty}}^{\boldsymbol{x}} \hat{p}_M(\boldsymbol{u}) d\boldsymbol{u}$ as the approximate CDF. Then*

$$\lim_{M \to \infty} \hat{F}_{\boldsymbol{C},M}(\boldsymbol{x}) = F_{\boldsymbol{C}}(\boldsymbol{x}).$$

Proof Sketch. By combining the error bounds in Lemma 7 and Lemma 8 and applying triangle inequality, we can show the end-to-end error is bounded by $|\hat{F}_{\boldsymbol{C},t}(\boldsymbol{x}) - F_{\boldsymbol{C}}(\boldsymbol{x})| \leq \frac{1}{M \cdot \hat{z} z}(\theta\mu z + U\mu F_{\boldsymbol{C}}(\boldsymbol{x}))$, where $\theta = \|\boldsymbol{x} - \boldsymbol{a}\|$ is the distance from $\boldsymbol{x}$ to $\boldsymbol{a} = [a_1, a_2, ..., a_N]$. Recall $\boldsymbol{C} = \bigotimes_{i=1}^{N}[a_i, b_i]$, so $\boldsymbol{a}$ is the lower bound of $\boldsymbol{C}$. Then θ, μ (Lipschitz constant of f), z (normalizing constant), U (volume of $\boldsymbol{C}$), and $F_{\boldsymbol{C}}(\boldsymbol{x})$ are all constants regarding M, and $\hat{z} \to z > 0$ as $M \to \infty$. Hence $|\hat{F}_{\boldsymbol{C},t}(\boldsymbol{x}) - F_{\boldsymbol{C}}(\boldsymbol{x})| \to 0$ as $M \to \infty$. □

We allow a user to provide a bounded domain $\boldsymbol{C}$, or infer it with automatically with a heuristic (Section 4). Although AQUA's formal guarantee is in a bounded domain, it can give runtime warnings when any prior or likelihood has probability

Algorithm 2 Posterior Interval Analysis with Adaptive Interval

```
1: procedure PosteriorAdaptiveAnalysis(Prog,M,t_0,t_dist)
2:     C ← GetInitBounds(Prog,t_0)                          ▷ C=[C_1,C_2,...,C_N]
3:     changed ← True
4:     while changed do                                     ▷ Stop if C no longer changes
5:         (p̂,Marginal) ← PosteriorAnalysis(Prog,M,C)       ▷ Apply analysis on C
6:         changed ← False
7:         for x_i ∈ SampledVars(Prog) do                   ▷ Adapt C_i for each variable
8:             p̂_i(x_i) ← Marginal[x_i]
9:             if ∃x_i ∈ C_i, p̂_i(x_i) < t_dist then
10:                C_i ← [inf{x_i|p̂_i(x_i) > t_dist},sup{x_i|p̂_i(x_i) > t_dist}]
11:                changed ← True
12:            end if
13:        end for
14:    end while
15: return (p̂, Marginal)
16: end procedure
```

greater than a given threshold on the rest of the domain $\mathbb{R}^N - \boldsymbol{C}$. If AQUA does not give any warning, the final error caused by truncating infinite domain into $\boldsymbol{C}$ will be smaller than the threshold.

4 AQUA Analysis Optimizations

Adaptive Intervals. To find the suitable bounded intervals $\boldsymbol{C}=[C_1,C_2,...,C_N]$ that cover most probability, we design a adaptive algorithm (Algorithm 2) to adjust $\boldsymbol{C}$ the based on the result from last run. Algorithm 2 takes as inputs the program, the vector of number of intervals, and two thresholds t_0 and t_{dist} for deciding the interval bounds $\boldsymbol{C}$. Increasing C_i or increasing the number of intervals in C_i will help reduce the approximation error.

The function *GetInitBounds* (Line 2) takes the prior distribution of each x_i as a rough estimate of the distribution to determine an initial interval split. If the domain of the prior distribution is bounded in $[a_i,b_i]$ where $-\infty < a_i < b_i < \infty$, e.g. `x_i ~ uniform(a,b)`, AQUA divides $[a_i,b_i]$ into M_i equi-length intervals, each with length $(b_i-a_i)/M_i$, where M_i is given in $\boldsymbol{M}$ by the user. If the distribution is not bounded, e.g. `x_i ~ normal(0,1)`, the user can specify a threshold t_0 for AQUA to infer C_is such that values from the prior being out of C_is has probability smaller than t_0. Otherwise by default we set $t_0 = 4 \cdot 10^{-32}$.

In each iteration, the algorithm applies the analysis on the current $\boldsymbol{C}$ (line 5) and check if we need to adapt $\boldsymbol{C}$. We adapt $\boldsymbol{C}$ when any variable x_i has density value $\hat{p}_i(x_i)$ being almost about 0 – smaller than the user provided threshold t_{dist} (e.g. 10^{-8}) (line 8-10). We shrink C_i to focus on the smallest area with density greater than a given threshold t_{dist}. With the same number of intervals M_i, the smaller C_i will produce thiner intervals and result in more accurate results. Practically, this adaptive algorithm is as accurate but is much more efficient than naively increasing

the number of intervals M_i on the whole initial domain C_i. Suppose the program takes A adaptive iterations, and it has N variables and each variable has the same number of intervals M, Algorithm 2 has the time complexity $\mathcal{O}(A \cdot N \cdot M^N)$ and the space complexity $\mathcal{O}(M^N)$. In our experiments, A is usually less than 5.

Improving Inference for Many Local Variables. In this optimization we change the analysis of statements in Section 3 to marginalize the local variables as soon as possible. Local variables are those defined and only used in local blocks (e.g. in for-loop and if-then-else from Figure 5).

By marginalizing out the local variables, we avoid repeatedly computing the joint density on the unused variables. For example, in a robust model one may naively calculate the joint density via $\hat{f}(x) = \prod_{i=1}^{D} \text{d\_pdf}(x, \text{w}_i)$, where w_is are local variables defined in each loop body. This requires keeping a $(D{+}1)$-dimensional density cube to capture all the variables x and w_is. Instead, our optimization divides the above product into calculating the individual $\text{d\_pdf}(x, \text{w}_i)$, when w_i leaves its scope, so we do not carry the current w_i to the next iteration. In each iteration we only operate on a 2-dimensional Density Cube for variables x and a single w_i. If out of N variables in the program D are local variables we will have a time complexity $\mathcal{O}(N \cdot M^{N-D})$ for Algorithm 1 (while the original is $\mathcal{O}(N \cdot M^N)$).

5 Methodology

We evaluate AQUA on 24 probabilistic programs collected from existing literature. We compare the execution time of AQUA on these programs with other probabilistic programming languages: Stan [5], PSI [8], and SPPL [19]. We implement AQUA in Java using ND4J library for tensor computation, and run all experiments on Intel Xeon 3.6 GHz machine with 6 cores and 32 GB RAM. For numerical stability, we use log probability/density (instead of original probability/density) for Density Cube.

Benchmarks. Table 2 presents the benchmarks obtained from the literature. Column **Description** summarizes the task of each program. Column **Distributions** shows the distributions of observable and latent variables. For example, the distributions in program "prior_mix" are one Bernoulli (B), one Mixture of two Normals ($N{+}N$), and 10 Student-T distributions (T^{10}). All posterior distributions are continuous. Column **#D** shows the number of data observations, **#N** shows the number of random variables in the program.

Comparing Posterior Distributions. The Kolmogorov-Smirnov (KS) statistic measures the distance between two probability distributions. We use the KS statistic for the accuracy evaluation in the analysis. Let F_{truth} and $\hat{F}$ denote the posterior distributions of the variable x from the original input data and the noisy data respectively, the *KS statistic* is defined as $KS = \sup_x \left| F_{truth}(x) - \hat{F}(x) \right|$, namely, the maximum difference in the cumulative distribution functions. The KS statistic takes a value between 0 (most close distributions) and 1 (most different distributions). Therefore, smaller KS statistic implies better accuracy.

Experimental Setup. We manually derived the *ground truth* posterior distributions for all the programs. We run AQUA with the adaptive algorithm described in Section 4. We use the equal number of $M = \max\{60, \lceil 40000^{(1/N)} \rceil\}$ intervals

Table 2: Program Description and Characteristics

	Description	Distributions	#D	#N
prior_mix	Mixture model[9]	$B\times(N+N)\times T^{10}$	10	1
zeroone	Bayesian neural network[3]	$U^2\times M^{20}$	20	2
tug	Causal cognition model[10]	$U^2\times(N+N)^2\times B^{40}$	40	2
altermu	Model with param symmetry[18]	$N^3\times N^{40}$	40	3
altermu2	Model with param symmetry[18]	$U^2\times N^{40}$	40	2
neural	Bayesian neural network[17]	$U^2\times(B\times M)^{39}$	39	2
normal_mixture	Mixture model with mixing rate[21]	$N^2\times Be\times(B\times(N+N))^{63}$	63	3
mix_asym_prior	Mixture model with scale params[21]	$N^2\times G^2\times(B\times(N+N))^{40}$	40	4
logistic	Logistic regression[21]	$U^2\times(B\times M)^{100}$	100	2
logistic_RW	Reweighted logistic regression[21,24]	$U^2\times Be^{100}\times(B\times M)^{100}$	100	102
anova	Linear regression [21]	$U^2\times N^{40}$	40	2
anova_RP	Localized linear regression[21,23]	$U^2\times G^{40}\times N^{40}$	40	42
anova_RW	Reweighted linear regression[21,24]	$U^2\times Be^{40}\times N^{40}$	40	42
lightspeed	Linear regression[21]	$N\times U\times N^{66}$	66	2
lightspeed_RP	Localized linear regression[21,23]	$N\times U\times G^{66}\times N^{66}$	66	68
lightspeed_RW	Reweighted linear regression[21,24]	$N\times U\times Be^{66}\times N^{66}$	66	68
unemployment	Linear regression[21]	$N^2\times U\times N^{40}$	40	3
unemployment_RP	Localized linear regression[21,23]	$N^2\times U\times G^{40}\times N^{40}$	40	43
unemployment_RW	Reweighted linear regression[21,24]	$N^2\times U\times Be^{40}\times N^{40}$	40	43
timeseries	Timeseries analysis[21]	$U^3\times N^{39}$	39	3
gammaTransform	Transformed param[19]	G	0	3
GPA	Hybrid continuous & discrete distr.[14]	$B\times(B\times(A+U)+B\times(A+U))$	1	3
radar_query1	Bayesian network in robotics[8]	$B\times(A+B)\times U\times N\times(Tr+Tr)$	2	6
radar_query2	Bayesian network in robotics[8]	$B\times(A+B)\times U^2\times N\times Tr$	1	6

Distributions: A: Atomic, B: Bernoulli, Be: Beta, G: Gamma, M: Softmax, N: Normal, T: Student-T, Tr: Triangular, U: Uniform. '+' represents the mixture of two distributions, and '×' represents the product of the individual density functions in the joint probability density function.

for each variable, where N is the number of sampled variables, so that the total number intervals $M^N \geq 40000$. Rounding up the total number of intervals to 40000 does not significantly affect time but will guarantee more accurate results. We test Stan on its two major inference algorithms, NUTS (a variant of MCMC) and ADVI (a variant of variational inference). For fair comparison, we allow running VI/NUTS until it reaches the same accuracy level (in KS distance) as AQUA and report the average time, or until it reaches the maximum iterations (fixed at 400000 for both VI and NUTS). We set the timeout to be 20 minutes for all the inference tools.

6 Evaluation

6.1 Runtime and Accuracy Comparison

Table 3 presents the runtime and accuracy comparison of AQUA with Stan, PSI, and SPPL. Column **Program** shows the name of the probabilistic program. Columns **Time (s)** show the execution time (in seconds) of each tool, averaged across 5 runs. We report the total time for computing joint density and marginals for all sampled variables. Columns **Error** show the error (KS distance, Section 5) of each tool vs. the ground truth when run for the same time, averaged across 5 runs.

Overall, **AQUA** (Column 2-3) solves the probabilistic programs with average time 5.08s, median time 1.35s. For 20 out of 24 programs, it takes less than two seconds to compute the results. AQUA results in average error 0.01, median error 0.01, and maximum error 0.02. With our optimization on local variables (Section 4), we are able to handle the 7 robust programs which have 42-102 variables, which might timeout with a naive approach.

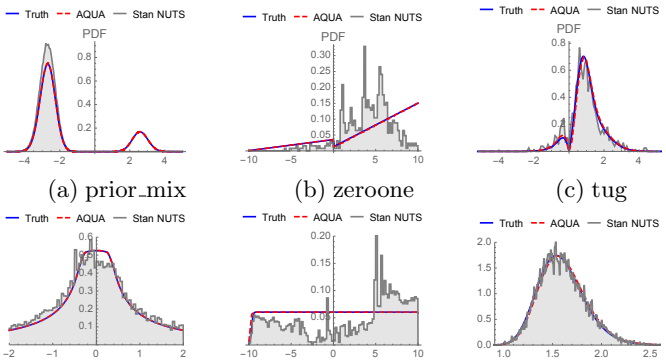
Table 3: Runtime Comparison for AQUA, Stan, PSI, and SPPL. Stan column shows time needed reach AQUA's accuracy.

Program	**AQUA**		**Stan VI**		**Stan NUTS**		**PSI**	**SPPL**
	Time(s)	Error	Time(s)	Error	Time(s)	Error	Time (s)	Time (s)
prior_mix	4.77	0.02	0.53	0.31	5.67	0.19	inte	⊘
zeroone	0.98	0.00	0.44	0.21	630.73	0.21	91.16	⊘
tug	0.83	0.01	1.20	0.25	519.94	0.06	inte	⊘
altermu	1.35	0.00	0.96	0.31	29.46	0.03	inte	⊘
altermu2	0.76	0.00	0.75	0.34	25.98	0.07	inte	⊘
neural	0.85	0.01	0.82	0.03	5.10	0.01	t.o.	⊘
normal_mixture	1.19	0.02	1.02	0.12	25.67	0.04	t.o.	⊘
mix_asym_prior	24.63	0.02	1.04	0.09	16.41	0.03	t.o.	⊘
logistic	0.99	0.02	0.74	0.07	17.31	0.02	t.o.	⊘
logistic_RW	1.87	0.01	15.37	0.09	72.45	0.02	t.o.	⊘
anova	0.90	0.01	0.75	0.07	6.72	0.02	inte	⊘
anova_RP	1.55	0.01	6.89	0.07	77.48	0.02	t.o.	⊘
anova_RW	1.40	0.01	6.93	0.06	24.67	0.02	t.o.	⊘
lightspeed	0.74	0.00	0.71	0.04	3.56	0.00	inte	⊘
lightspeed_RP	1.37	0.01	6.18	0.06	61.37	0.02	t.o.	⊘
lightspeed_RW	1.09	0.02	6.19	0.05	61.37	0.05	t.o.	⊘
unemployment	1.44	0.02	0.64	0.21	5.07	0.01	inte	⊘
unemployment_RP	42.34	0.01	6.78	0.25	12.46	0.01	t.o.	⊘
unemployment_RW	27.41	0.02	7.07	0.23	2.53	0.01	t.o.	⊘
timeseries	1.55	0.01	0.87	0.23	12.66	0.01	inte	⊘
gammaTransform	0.72	0.00	0.62	0.05	3.01	0.01	inte	1.30
GPA	0.46	0.02	⊘	⊘	⊘	⊘	0.12	0.05
radar_query1	0.87	0.01	⊘	⊘	⊘	⊘	inte	⊘
radar_query2	1.82	0.02	⊘	⊘	⊘	⊘	inte	⊘
Avg	5.08	0.01	3.17	0.15	77.12	0.04	⊘	⊘
Median	1.35	0.01	0.99	0.10	20.99	0.02	⊘	⊘

[*time*] : VI or NUTS takes more time than AQUA, or AQUA take more time than VI and NUTS.
[*error*] Has the error (in terms of a KS distance) larger than 0.01 from the best solution.
"⊘": the PPL cannot work on the program. "t.o.": timeout, "inte": evaluates to unsolved integrals.

Stan VI (Column 4-5) finishes fast but results in significantly larger error than AQUA or Stan NUTS. The average error from VI is 0.15, minimum error is 0.03 and maximum error is 0.34. For all cases, VI cannot reach the same accuracy level as AQUA. While VI often fits the posterior means correctly but it is not able to capture the joint distribution shape especially when it is non-Gaussian (it is a well known characteristic of VI). **Stan NUTS** (Column 6-7) takes more time than AQUA to reach the same level of accuracy of AQUA, although in theory NUTS will converge to the true distribution with enough iterations. AQUA provides the similar (with difference <0.01) or even better accuracy (with smaller KS distance) in all cases for NUTS and NUTS fails to reach the same accuracy level by the maximum number of iterations in 12 cases.

PSI (Column 8) and **SPPL** (Column 9) are not able to give result in many cases. PSI does not finish running within 20 minutes in 11 cases, or evaluates to unsolved integrals in 11 cases, since the exact integration in posterior calculation is often intractable. SPPL does not allow transformed variables in `factor` statements, which is essential to specify the likelihood of the variables given observed data, and thus is inapplicable to most of the programs.



(a) prior_mix (b) zeroone (c) tug

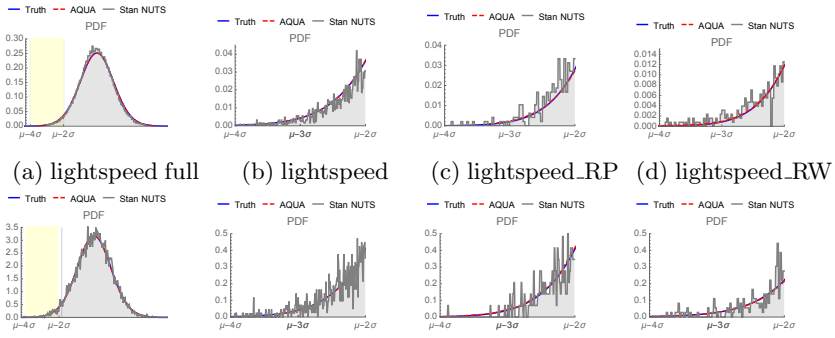
(d) altermu (e) altermu2 (f) anova_RP

Fig. 6: Programs handled by AQUA for which Stan NUTS is imprecise.

Figure 6 presents the posterior densities from six programs where Stan NUTS was not able to reach the same accuracy level of AQUA, within maximum iterations. X-axis shows the value of a variable in the program, Y-axis shows the posterior probability density of the variable. A solid blue line shows the ground truth, a dashed red line shows the density function computed from AQUA, the gray histogram shows the density estimated with samples from Stan NUTS after running for the same time as AQUA. For each program we present the posterior from one variable (the first one in alphabetical order); the posteriors from other variables show a similar pattern. These examples show that AQUA is able to closely track the density of mixture models with large difference in densities (**prior_mix**), non-differentiable distributions (**zeroone** and **tug**), models with variable symmetries (in **altermu** and **altermu2** such symmetries can cause non-identifiability of variables from data), and some robust models with strong correlation between variables that can form complicated posterior geometries (**anova_RP**).

6.2 Estimating the Tails of Posterior Distribution

We illustrate AQUA's ability to capture tails on several robust models. The distribution for robust models are often more spread-out than the original model, as



(a) lightspeed full (b) lightspeed (c) lightspeed_RP (d) lightspeed_RW

(e) anova full (f) anova (g) anova_RP (h) anova_RW

Fig. 7: Capturing tails by AQUA and Stan NUTS

they are designed to capture outliers in the data. We consider two different robust models: (1) Reparameterized-Localization (RP) [23], which assumes that each data point is from its distribution with a local variance variable; (2) Reweighting (RW) [24], which down-weights potential outliers in the data. We show the results from AQUA and NUTS running for the same amount of time, together with the ground truth. We omit VI since its accuracy is significantly worse.

Figure 7 presents the comparison of AQUA and NUTS. Plots (a),(e) are the full posterior distributions of original distribution. We highlight the left tail $[\mu-4\sigma, \mu-2\sigma]$, where μ is the posterior mean of and σ its standard deviation. Plots (b),(f) show the magnified tails from original distribution, plots (c),(g) show the tails from the RP transformation, and (d),(h) show the tails from RW transformation. AQUA is able to capture the tails precisely for both original and robust models, while Stan NUTS is less precise on the robust models (e.g., its KS metric is 0.05 compared to AQUA's 0.02).

7 Related Work

Symbolic Inference. Researchers have proposed several symbolic inference techniques in recent years [8,19,16,12]. Each of these techniques have limitations which AQUA improves upon. DICE [12] only supports discrete distributions. Hakaru [16] and PSI [8], which do exact inference using computer algebra, often cannot solve integrals for complicated probabilistic programs with continuous distributions (as our evaluation also shows for PSI). SPPL [19] does not allow users to specify the likelihood on transformed variables with continuous distributions. QCoral [4] and SYMPAIS [15] combine symbolic execution with sampling to solve the satisfaction probability of constraints, but they do not output the whole posterior. In contrast, AQUA supports a wide range of probabilistic models with continuous distribution, involving transformed or correlated random variables, and provides scalable, exact (or near exact), and interpretable solutions.

Volume Computation. Several works use volume computation methods to make a precise approximation of probabilistic inference [20,2,22]. These approaches have constraints on the form of programs they support, regarding conditioning and continuous distributions. For instance, Sweet et al. [22] support only discrete and FairSquare [2] approximates Gaussians with only five intervals; FairSquare [2] and Sankaranarayanan et al. [20] compute only the probability of an event, not the entire posterior. None of these systems can support conditioning on continuous variables, and thus we have not used them in our evaluation.

8 Conclusion

AQUA is a new inference algorithm which works on general, real-world probabilistic programs with continuous distributions. By using quantization with symbolic inference, AQUA solved all benchmarks in less than 43s (median 1.35s). Our evaluation shows that AQUA is more accurate than approximate algorithms and supports programs that are out of reach of state-of-the-art exact inference tools.

Acknowledgements. This research was supported in part by NSF Grants No. CCF-1846354, CCF-1956374, CCF-2008883, and Facebook PhD Fellowship.

References

1. Aguirre, A., Barthe, G., Hsu, J., Kaminski, B.L., Katoen, J.P., Matheja, C.: A pre-expectation calculus for probabilistic sensitivity. POPL (2021)
2. Albarghouthi, A., D'Antoni, L., Drews, S., Nori, A.: Fairsquare: probabilistic verification of program fairness (OOPSLA) (2017)
3. Bissiri, P., Holmes, C., Walker, S.: A general framework for updating belief distributions. Journal of the Royal Stat. Soc. Series B, Statistical Methodology **78**(5), 1103 (2016)
4. Borges, M., Filieri, A., d'Amorim, M., Păsăreanu, C.S., Visser, W.: Compositional solution space quantification for probabilistic software analysis. PLDI (2014)
5. Carpenter, B., Gelman, A., Hoffman, M., Lee, D., et al.: Stan: A probabilistic programming language. JSTATSOFT **20**(2) (2016)
6. Dutta, S., Legunsen, O., Huang, Z., Misailovic, S.: Testing probabilistic programming systems. In: FSE (2018)
7. Dutta, S., Zhang, W., Huang, Z., Misailovic, S.: Storm: program reduction for testing and debugging probabilistic programming systems. In: FSE (2019)
8. Gehr, T., Misailovic, S., Vechev, M.: PSI: Exact symbolic inference for probabilistic programs. CAV (2016)
9. Gelman, A., Stern, H.S., Carlin, J.B., Dunson, D.B., Vehtari, A., Rubin, D.B.: Bayesian data analysis. Chapman and Hall/CRC (2013)
10. Goodman, N., Tenenbaum, J.: Probabilistic Models of Cognition. `probmods.org`
11. Gorinova, M.I., Gordon, A.D., Sutton, C.: Probabilistic programming with densities in SlicStan: efficient, flexible, and deterministic. POPL (2019)
12. Holtzen, S., Van den Broeck, G., Millstein, T.: Scaling exact inference for discrete probabilistic programs. OOPSLA (2020)
13. Huang, Z., Wang, Z., Misailovic, S.: PSense: automatic sensitivity analysis for probabilistic programs. ATVA (2018)
14. Laurel, J., Misailovic, S.: Continualization of probabilistic programs with correction. ESOP (2020)
15. Luo, Y., Filieri, A., Zhou, Y.: Sympais: Symbolic parallel adaptive importance sampling for probabilistic program analysis. arXiv preprint arXiv:2010.05050 (2020)
16. Narayanan, P., Carette, J., Romano, W., Shan, C.c., Zinkov, R.: Probabilistic inference by program transformation in hakaru (system description). FLOPS (2016)
17. Neal, R.M.: Bayesian learning for neural networks. Springer (2012)
18. Nishihara, R., Minka, T., Tarlow, D.: Detecting parameter symmetries in probabilistic models. arXiv preprint arXiv:1312.5386 (2013)
19. Saad, F.A., Rinard, M.C., Mansinghka, V.K.: SPPL: a probabilistic programming system with exact and scalable symbolic inference. PLDI (2021)
20. Sankaranarayanan, S., Chakarov, A., Gulwani, S.: Static analysis for probabilistic programs: Inferring whole program properties from finitely many paths. PLDI (2013)
21. (2018), `https://github.com/stan-dev/example-models`
22. Sweet, I., Trilla, J.M.C., Scherrer, C., Hicks, M., Magill, S.: Whats the over/under? probabilistic bounds on information leakage. POST (2018)
23. Wang, C., Blei, D.M.: A general method for robust bayesian modeling. Bayesian Analysis **13**(4), 1159–1187 (2018)
24. Wang, Y., Kucukelbir, A., Blei, D.M.: Robust probabilistic modeling with bayesian data reweighting. ICML (2017)