09-12 Nov
CAEPIA'15
Albacete

Parallel Importance Sampling in Conditional Linear Gaussian Networks

Antonio Salmerón, **Darío Ramos-López**, Hanen Borchani,
Antonio Fernández, Ana M. Martínez, Andrés Masegosa, Helge Langseth,
Anders Madsen and Thomas D. Nielsen

The AMIDST project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no 619209.

Outline of the talk

- Introduction
  - Probabilistic Inference
  - Conditional Linear Gaussian (CLG) Bayesian Networks
- Approximate Inference in CLG Bayesian Networks
  - Variational Inference
  - **Importance Sampling**
- Experimental results of **Parallel Importance Sampling**
- Conclusions

Probabilistic Inference

A Bayesian Network (BN) is a Directed Acyclic Graph (DAG) of variables $\mathbf{X} = \{X_1, \ldots, X_N\}$ with a joint probability distribution:

$$p(\mathbf{X}) = \prod_{i=1}^{N} p(X_i|Pa(X_i))$$

where $Pa(X_i)$ stands for the set of parents of X_i

$\mathbf{X}_E \subset \mathbf{X}$ Observed variables

$\mathbf{X}_I \subset \mathbf{X} \backslash \mathbf{X}_E$ Variables of interest

$$p(x_i|\mathbf{x}_E) = \frac{p(x_i, \mathbf{x}_E)}{p(\mathbf{x}_E)} = \frac{\sum_{\mathbf{x}_D \in \Omega_{\mathbf{X}_D}} \int_{\mathbf{x}_C \in \Omega_{\mathbf{X}_C}} p(\mathbf{x}, \mathbf{x}_E) \mathrm{d}\mathbf{x}_C}{\sum_{\mathbf{x}_{D_i} \in \Omega_{\mathbf{X}_{D_i}}} \int_{\mathbf{x}_{C_i} \in \Omega_{\mathbf{X}_{C_i}}} p(\mathbf{x}, \mathbf{x}_E) \mathrm{d}\mathbf{x}_{C_i}}$$

Probabilistic Queries

For continuous variables

$$p(a < X_i < b|\mathbf{x}_E) = \frac{\int_a^b \left(\sum_{\mathbf{x}_D \in \Omega_{\mathbf{x}_D}} \int_{\mathbf{x}_C \in \Omega_{\mathbf{x}_C}} p(\mathbf{x}, \mathbf{x}_E) \mathrm{d}\mathbf{x}_C \right) \mathrm{d}x_i}{\sum_{\mathbf{x}_{D_i} \in \Omega_{\mathbf{x}_{D_i}}} \int_{\mathbf{x}_{C_i} \in \Omega_{\mathbf{x}_{C_i}}} p(\mathbf{x}, \mathbf{x}_E) \mathrm{d}\mathbf{x}_{C_i}}$$

Or discrete variables

$$p(X_i = x_i|\mathbf{x}_E) = \frac{\sum_{\mathbf{x}_D \in \Omega_{\mathbf{x}_D}} \int_{\mathbf{x}_C \in \Omega_{\mathbf{x}_C}} p^{R(X_i = x_i)}(\mathbf{x}, \mathbf{x}_E) \mathrm{d}\mathbf{x}_C}{\sum_{\mathbf{x}_{D_i} \in \Omega_{\mathbf{x}_{D_i}}} \int_{\mathbf{x}_{C_i} \in \Omega_{\mathbf{x}_{C_i}}} p(\mathbf{x}, \mathbf{x}_E) \mathrm{d}\mathbf{x}_{C_i}}$$

Conditional Linear Gaussian (CLG) BN

A Conditional Linear Gaussian (CLG) network is a hybrid Bayesian network where

- The conditional distribution of each discrete variable X_D given its parents is a multinomial
- The conditional distribution of each continuous variable Z with discrete parents $\mathbf{X}_D$ and continuous parents $\mathbf{X}_C$, is

$$p(z|\mathbf{X}_D = \mathbf{x}_D, \mathbf{X}_C = \mathbf{x}_C) = \mathcal{N}(z; \alpha(\mathbf{x}_D) + \beta(\mathbf{x}_D)^\top \mathbf{x}_C, \sigma(\mathbf{x}_D))$$

for all $\mathbf{x}_D$ and $\mathbf{x}_C$, where α and β are the coefficients of a linear regression model of Z given $\mathbf{X}_C$, potentially different for each configuration of $\mathbf{X}_D$.

Inference in CLG Bayesian Networks

Exact inference requires a strong junction tree

→ **Computationally expensive**

Not suitable for dealing with streaming data or large networks

Approximate inference algorithms with a quick response are needed, such as:

- Variational Inference
- **Importance Sampling**

Variational Inference

It is a deterministic approximate inference technique, with iterative optimization of a variational approximation to the posterior distribution of interest.

For a posterior distribution of interest $p(\mathbf{x}_I|\mathbf{X}_E = \mathbf{x}_E)$ and a set $\mathcal{Q}$ of possible approximations, the variational approximation is defined as:

$$q^*_{\mathbf{x}_E}(\mathbf{x}_I) = \arg\min_{q \in \mathcal{Q}} D(q(\mathbf{x}_I)||p(\mathbf{x}_I|\mathbf{X}_E = \mathbf{x}_E))$$

$D(q||p)$ is the KL divergence between q and p

Variational Inference

A common approach is to employ a variational mean-field approximation of the posterior distribution

$$q^*_{\mathbf{x}_E}(\mathbf{x}_I) = \prod_{i \in I} q^*_{\mathbf{x}_E}(x_i)$$

During the optimization, the individual variational distributions are iteratively updated, holding the others fixed.

Updating a variational distribution essentially involves calculating the variational expectation of the logarithm of the original conditional distributions of the model.

This can be done efficiently and in closed form when the distributions involved are conjugate-exponential, using the general architecture of *variational message passing* (VMP)

Importance Sampling

Importance sampling is a versatile simulation technique. In case of inference in BNs amounts to transforming the numerator of

$$p(a < X_i < b|\mathbf{x}_E) = \frac{\int_a^b \left(\sum_{\mathbf{x}_D \in \Omega_{\mathbf{X}_D}} \int_{\mathbf{x}_C \in \Omega_{\mathbf{X}_C}} p(\mathbf{x}, \mathbf{x}_E) \mathrm{d}\mathbf{x}_C \right) \mathrm{d}x_i}{\sum_{\mathbf{x}_{D_i} \in \Omega_{\mathbf{X}_{D_i}}} \int_{\mathbf{x}_{C_i} \in \Omega_{\mathbf{X}_{C_i}}} p(\mathbf{x}, \mathbf{x}_E) \mathrm{d}\mathbf{x}_{C_i}}$$

by multiplying and dividing by a distribution which is easier to handle and from which samples can easily be drawn.

Importance Sampling

Let θ denote the numerator of

$$p(a < X_i < b|\mathbf{x}_E) = \frac{\int_a^b \left(\sum_{\mathbf{x}_D \in \Omega_{\mathbf{X}_D}} \int_{\mathbf{x}_C \in \Omega_{\mathbf{X}_C}} p(\mathbf{x}, \mathbf{x}_E) \mathrm{d}\mathbf{x}_C \right) \mathrm{d}x_i}{\sum_{\mathbf{x}_{D_i} \in \Omega_{\mathbf{X}_{D_i}}} \int_{\mathbf{x}_{C_i} \in \Omega_{\mathbf{X}_{C_i}}} p(\mathbf{x}, \mathbf{x}_E) \mathrm{d}\mathbf{x}_{C_i}}$$

$$\theta = \int_a^b h(x_i) \mathrm{d}x_i \qquad h(x_i) = \sum_{\mathbf{x}_D \in \Omega_{\mathbf{X}_D}} \int_{\mathbf{x}_C \in \Omega_{\mathbf{X}_C}} p(\mathbf{x}, \mathbf{x}_E) \mathrm{d}\mathbf{x}_C$$

$$\theta = \int_a^b h(x_i) \mathrm{d}x_i = \int_a^b \frac{h(x_i)}{p^*(x_i)} p^*(x_i) = E_{p^*}\left[\frac{h(X_i^*)}{p^*(X_i^*)} \right]$$

p^* is a probability density function on (a, b) called the *sampling distribution*

X_i^* is a random variable with density p^*

Importance Sampling

Let $X_i^{*(1)}, \ldots, X_i^{*(m)}$ be a sample drawn from p^*. Then

$$\hat{\theta}_1 = \frac{1}{m} \sum_{j=1}^{m} \frac{h(X_i^{*(j)})}{p^*(X_i^{*(j)})} \quad \text{is an unbiased estimator of } \theta$$

$$\begin{aligned} \text{Var}(\hat{\theta}_1) &= \text{Var}\left(\frac{1}{m} \sum_{j=1}^{m} \frac{h(X_i^{*(j)})}{p^*(X_i^{*(j)})} \right) = \frac{1}{m^2} \sum_{j=1}^{m} \text{Var}\left(\frac{h(X_i^{*(j)})}{p^*(X_i^{*(j)})} \right) \\ &= \frac{1}{m^2} m \text{Var}\left(\frac{h(X_i^*)}{p^*(X_i^*)} \right) = \frac{1}{m} \text{Var}\left(\frac{h(X_i^*)}{p^*(X_i^*)} \right). \end{aligned}$$

Importance Sampling – EW algorithm

Evidence Weighting (EW) is a procedure for selecting the sampling distribution

Each variable is sampled from its conditional density given its parents.
Sampling from parents to children in the network.

Observed variables are not sampled, but instantiated to the observed value.

Then, h involves the product of all the conditional, while p^* involves the same distributions except those corresponding to observed variables.

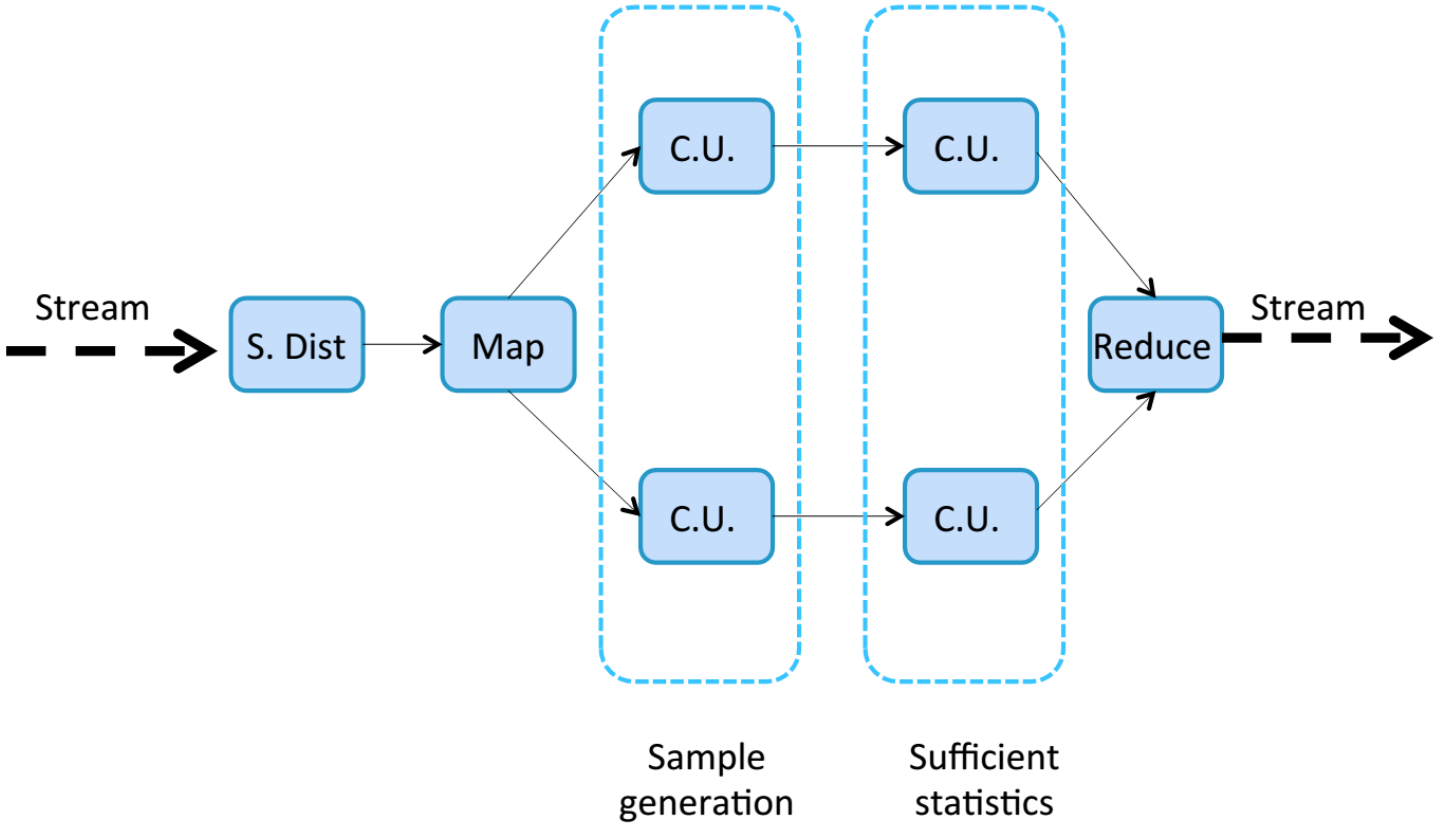
The probability of evidence has to be estimated as well. Thus, instead of taking a sampling distribution defined on (*a,b*) it must be defined on the entire range of the variable.

Importance Sampling – EW algorithm

Function EW($\mathbf{X}$,P,$\mathbf{x}_E$,X,a,b,M)
Input: The set of variables in the network, $\mathbf{X} = \{X_1, \ldots, X_N\}$ in topological order. The distributions in the network $P = \{p_1, \ldots, p_N\}$. Evidence $\mathbf{X}_E = \mathbf{x}_E$. The target variable X. Sample size M.
Output: An estimation of $P(a < X < b | \mathbf{X}_E = \mathbf{x}_E)$
begin
  Initialization:
  $s_1 \leftarrow 0$; $s_2 \leftarrow 0$.
  for $j \leftarrow 1$ **to** M **do**
    Sample generation:
    $w_1 \leftarrow 1$; $w_2 \leftarrow 1$.
    for $i \leftarrow 1$ **to** N **do**
      if $X_i \notin \mathbf{X}_E$ **then**
        Simulate a value $x_i^{(j)}$ for X_i using $p_i(x_i|Pa(x_i))$.
        $w_2 \leftarrow w_2 * p_i(x_i^{(j)}|Pa(x_i))$.
      end
      else
        Let $x_i^{(j)}$ be the value of X_i in $\mathbf{X}_E$.
      end
      $w_1 \leftarrow w_1 * p_i(x_i^{(j)}|Pa(x_i))$.
    end
    if $w_1 \neq 0$ **then**
      Let $x^{(j)}$ be the value of X in the simulated configuration $x_1^{(j)}, \ldots, x_N^{(j)}$.
      if $x^{(j)} \in (a, b)$ **then**
        $s_1 \leftarrow s_1 + w_1/w_2$
      end
      $s_2 \leftarrow s_2 + w_1/w_2$
    end
  end
  return s_1/s_2 .
end

Algorithm 1: The EW algorithm for answering a probabilistic query.

Importance Sampling – Parallelization



Experimental Results

Experiment set-up:

- Two randomly generated CLG BN of 10 and 500 variables, with double of links than nodes on each network.
- Observations in 5% of variables, randomly selected.
- Random queries of the form $p(a < X_i < b|\mathbf{x}_E)$ with $b - a = 1$.
- Each query answered with VPM and EW, with sample sizes of 1000, 5000, 10000.
- Each experiment replicated with 1, 2, 4, 8, 12, 16 and 20 cores, on a dual-processor AMD Opteron 2.8GHz server with 32 cores and 64GB of RAM, running Linux Ubuntu 14.04.1 LTS.
- Each run repeated 10 times to obtain the averaged runtime and the error given by the χ^2 divergence $\chi^2 = \frac{1}{n}\sum_{i=1}^{10}\frac{(q_i - p_i)^2}{p_i}$

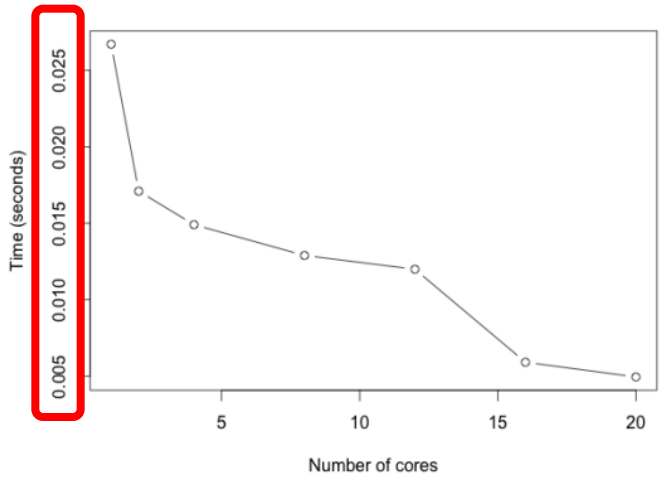
Parallel Importance Sampling in Conditional Linear Gaussian Networks

Experimental Results – Run time

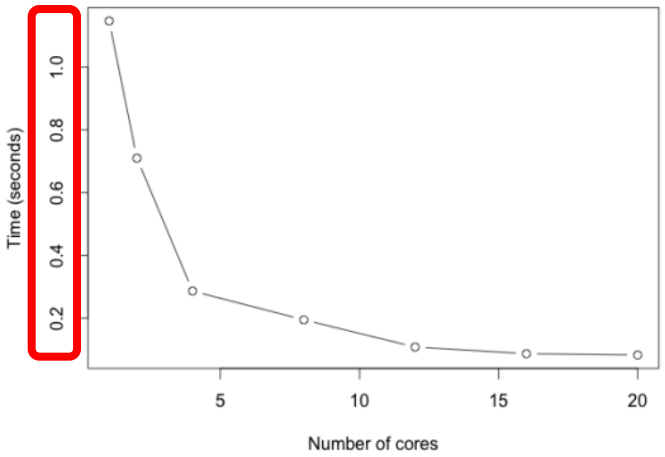
	10 vars.	500 vars.
Run time (seconds)	0.0739	9.6917
Error	0.4657	2.2759

Table 1. Error and run times for VMP

10 variables



500 variables



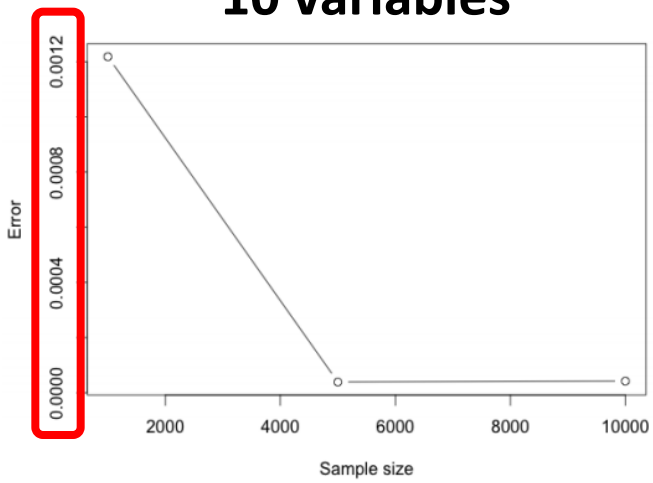
Parallel Importance Sampling in Conditional Linear Gaussian Networks

Experimental Results - Accuracy

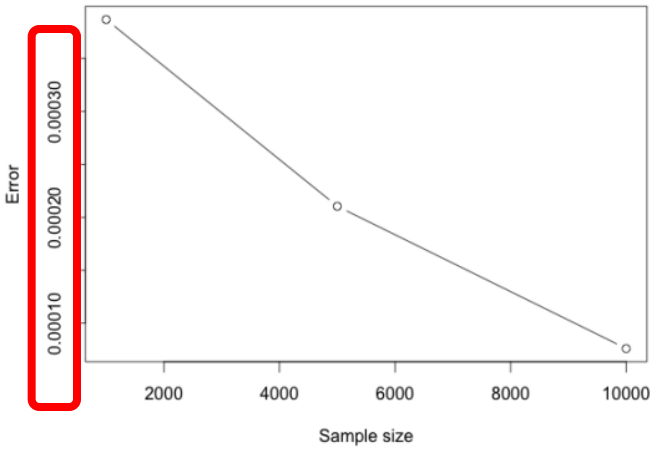
	10 vars.	500 vars.
Run time (seconds)	0.0739	9.6917
Error	0.4657	2.2759

Table 1. Error and run times for VMP

10 variables



500 variables



Conclusions

Two general approaches for probabilistic inference in CLG networks have been tested. According to the experiments:

- The Parallel Importance Sampling method (EW) outperforms VMP in both speed and precision.
- The quick responses provided by EW suggest that it might be an appropriate inference method for streaming evidence.
- However, the experiments were limited and should be extended to:
  - Larger networks with more links.
  - Distributions with a high concentration of extreme probabilities.

Thank you for your attention

Analysis of Massive Data Streams

http://www.amidst.eu