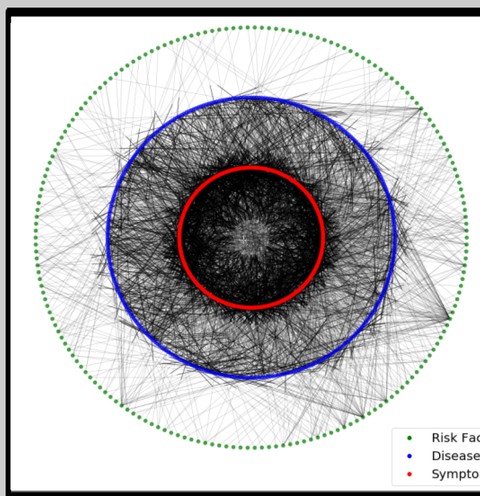
# A Universal Marginalizer for Amortized Inference in Generative Models

Laura Douglas[*†], Iliyan Zarov[*†], Konstantinos Gourgoulias[†], Chris Lucas[†], Chris Hart[†], Adam Baker[†], Maneesh Sahani[‡], Yura Perov[†], Saurabh Johri[†]

† babylon, London, UK. ‡ Gatsby Computational Neuroscience Unit, University College London. * equal contribution

## Abstract

We consider the problem of inference in a causal generative model where the set of available observations differs between data instances. We show how combining samples drawn from the graphical model with an appropriate masking function makes it possible to train a single neural network to approximate all the corresponding conditional marginal distributions and thus amortize the cost of inference. We further demonstrate that the efficiency of importance sampling may be improved by basing proposals on the output of the neural network. We also outline how the same network can be used to generate samples from an approximate joint posterior via a chain decomposition of the graph.



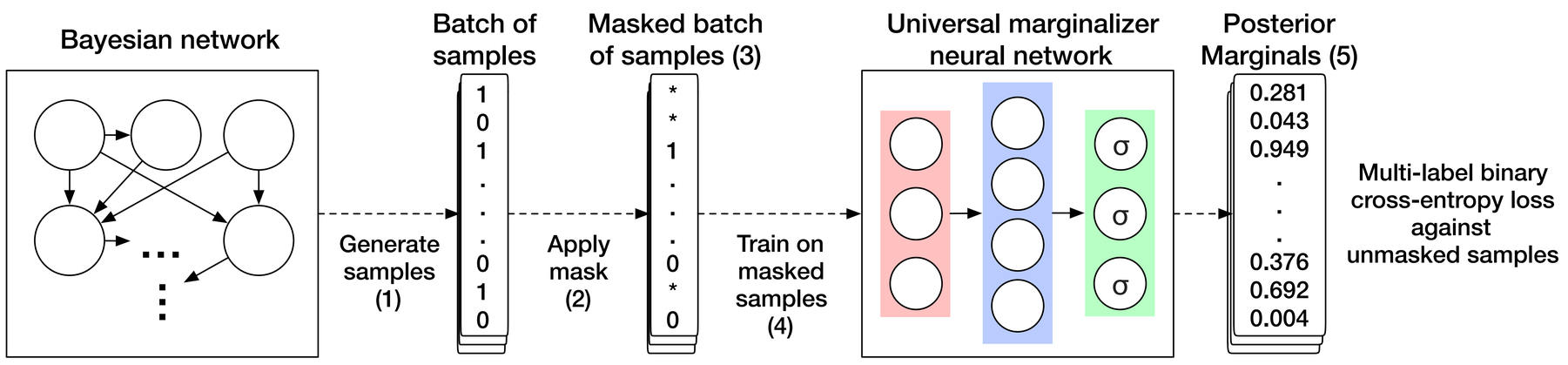

## Motivation

At babylon we have built a Bayesian network (BN) which can be used to diagnose diseases from user answers about symptoms and risk factors. The network is large and the nodes are in three layers and heavily connected, as visualised to the left. This means traditional inference methods are either slow or inaccurate. The flow of questions is dependent on the outcomes of inference on this BN and so speeding up inference reduces latency for the user. Furthermore, faster inference results in huge computational savings. These reasons motivate our amortized inference approach, which becomes more beneficial as the number of our daily users increases.

## Universal Marginalizer

A function approximator, such as a single neural network, trained to return all posterior marginal distributions of a Bayesian network given any set of observations.

**1. Generating Data:** The UM can be trained off-line by generating unbiased samples from the BN using ancestral sampling. Each sample is a binary vector which are the values the classifier will learn to predict.




**2. Masking:** To represent a state where much of the graph is unobserved, a subset of the nodes in the sample must be hidden, or masked. We choose to probabilistically mask a sample in an unbiased way by defining a masking probability, which is applied to each node in the sample.

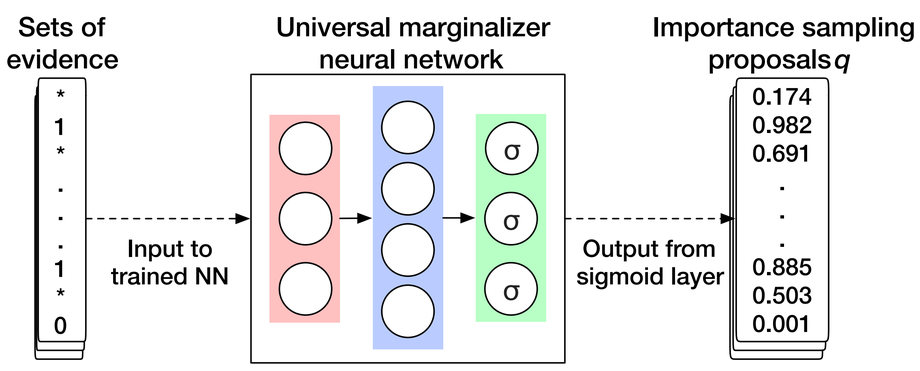
**3. Representation of the Unobserved/Masked Nodes:** We experimented with different representations and found the best to be a binary variable to represent if it was observed and a float to represent its prior probability (so 1 or 0 if observed and the prior if not).

**4. Training a neural net with Cross Entropy Loss:** We train the network using a binary cross entropy loss function in a multi-label classification setting to predict the state of all observed and unobserved nodes.

**5. Outputs: Posterior Marginals:** The desired posterior marginals are the output of the sigmoid layer. This result can already be used as a rough posterior estimate, however results can be further improved by combining with Importance Sampling.

## Generating Proposals Using the UM

To obtain a guarantee of asymptotic unbiasedness, while making use of the speed of the approximate solution, we use the UM for proposals in importance sampling.




A naive approach might be to sample each node independently from their respective marginal distribution, parameterized by each element in q above. However, the product of the (approximate) posterior marginals may be very different to the true posterior joint, even if the marginal approximations are good. This can cause huge variance in the Importance Sampling weights. We propose two methods to approximate the true joint: Sequential UM and Hybrid UM-Likelihood.

## Sequential UM

In SUM-IS we sequentially approximate the conditional marginal for a node i given the current sampled state and evidence, to get the optimal proposal.

$$Q_i^\star = P(X_i|\{X_1,\ldots,X_{i-1}\}\cup\mathbf{X}_\mathcal{O}) \approx \mathrm{UM}(\tilde{\mathbf{x}}_{\mathcal{S}\cup\mathcal{O}})_i = Q_i.$$

The full sample is thus drawn from an implicit encoding by the UM of the (approximate) posterior joint distribution, as can be seen by observing the product of sample probabilities:

$$Q = \mathrm{UM}(\tilde{\mathbf{x}}_\mathcal{O})_1 \prod_{i=2}^{N} \mathrm{UM}(\tilde{\mathbf{x}}_{\mathcal{S}\cup\mathcal{O}})_i \approx P(X_1|\mathbf{X}_\mathcal{O}) \prod_{i=2}^{N} P(X_i|X_1,\ldots,X_{i-1},\mathbf{X}_\mathcal{O}).$$

**Algorithm 1** Sequential Universal Marginalizer importance sampling

1: Order the nodes topologically $X_1,\ldots X_N$, where $N$ is the total number of nodes.
2: **for** $j$ in [1,...,$M$] (where $M$ is the total number of samples): **do**
3: $\quad\tilde{\mathbf{x}}_\mathcal{S} = \emptyset$
4: $\quad$**for** $i$ in [1,...$N$]: **do**
5: $\quad\quad$sample node $x_i$ from $Q(X_i) = \mathrm{UM}(\tilde{\mathbf{x}}_{\mathcal{S}\cup\mathcal{O}})_i \approx P(X_i|\mathbf{X}_\mathcal{S},\mathbf{X}_\mathcal{O})$
6: $\quad\quad$add $x_i$ to $\tilde{\mathbf{x}}_\mathcal{S}$
7: $\quad[\mathbf{x}_\mathcal{S}]_j = \tilde{\mathbf{x}}_\mathcal{S}$
8: $\quad w_j = \prod_{i=1}^{N} \frac{P_i}{Q_i}$ (where $P_i$ is the likelihood, $P_i = P(X_i = x_i|\mathbf{x}_{\mathcal{S}\cap\mathrm{Pa}(X_i)})$ and $Q_i = Q(X_i = x_i)$)
9: $E_p[X] = \frac{\sum_{j=1}^{M} X_j w_j}{\sum_{j=1}^{M} w_j}$ (as in standard IS)

## Hybrid UM-Likelihood Proposals

The full SUM-IS process requires sequential sampling and many evaluations of the UM, which may be costly. However, in Hybrid UM-Likelihood, a single UM output of all marginals may be combined with ancestral sampling, where nodes are sampled in topological order.

$$Q(X_i) = \beta \cdot \mathrm{UM}(\tilde{\mathbf{x}}_\mathcal{O})_i + (1-\beta)\cdot P(X_i|\mathbf{x}_{\mathcal{S}\cap\mathrm{Pa}(X_i)}).$$

approx posterior (conditioned on evidence))

prior conditioned on parents - whether previously sampled or evidence

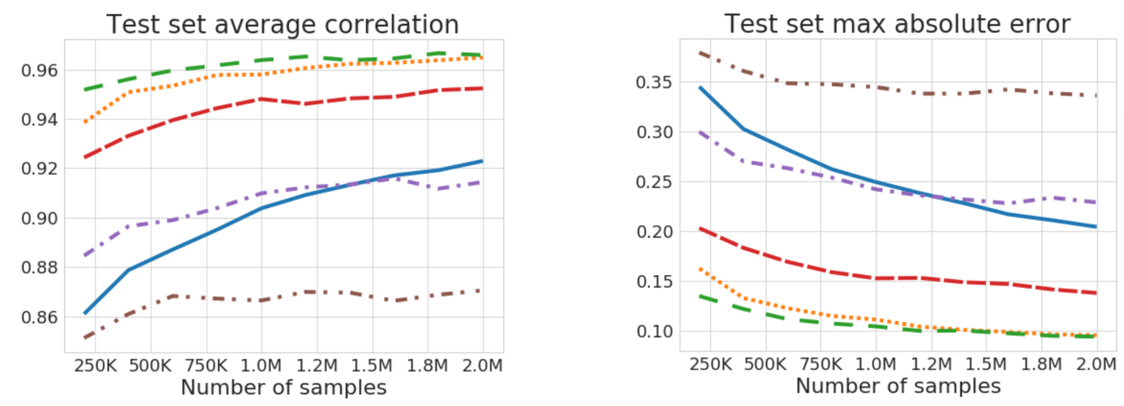
Here, each node in the proposal is drawn either from the UM approximate marginal given the observed evidence, independently of previously sampled nodes, or according to its prior dependence on previously sampled nodes (and any ancestral evidence), independently of evidence nodes that fall later in the topological sequence. This approach expects to blend these two forms of dependence, generating a reasonable IS proposal.

**Algorithm 2** Hybrid UM-IS

1: Order the nodes topologically $X_1,\ldots X_N$, where $N$ is the total number of nodes.
2: **for** $j$ in [1,...,$M$] (where $M$ is the total number of samples): **do**
3: $\quad\tilde{\mathbf{x}}_\mathcal{S} = \emptyset$
4: $\quad$**for** $i$ in [1,...$N$]: **do**
5: $\quad\quad$sample node $x_i$ from $Q(X_i) = \beta\mathrm{UM}(\tilde{\mathbf{x}}_\mathcal{O})_i + (1-\beta)P(X_i = x_i|\mathbf{x}_{\mathcal{S}\cap\mathrm{Pa}(X_i)})$
6: $\quad\quad$add $x_i$ to $\tilde{\mathbf{x}}_\mathcal{S}$
7: $\quad[\mathbf{x}_\mathcal{S}]_j = \tilde{\mathbf{x}}_\mathcal{S}$
8: $\quad w_j = \prod_{i=1}^{N} \frac{P_i}{Q_i}$ (where $P_i$ is the likelihood, $P_i = P(X_i = x_i|\mathbf{x}_{\mathcal{S}\cap\mathrm{Pa}(X_i)})$ and $Q_i = Q(X_i = x_i)$)
9: $E_p[X] = \frac{\sum_{j=1}^{M} X_j w_j}{\sum_{j=1}^{M} w_j}$ (as in standard IS)

## Results

We observe a reduced max error and an increased ESS for small values of beta. Standard ancestral sampling (beta=0, blue line) reaches 92% correlation after 2 million samples, whereas hybrid proposals with beta=0.25 (green line) exceeds 95% after only 250,000 samples, ultimately achieving 96% correlation in 2 million samples. Using hybrid UM-IS we achieve higher accuracy, lower variance and thus a significant reduction in computational cost per inference.




Figures: Hybrid importance sampling performance for various values of the mixing parameter beta between pure ancestral sampling proposals (beta = 0) and UM marginals independent of the sampled state of unobserved nodes (beta = 1).

As beta grows towards 1 the variance of the weights increases beyond that of standard ancestral sampling. The results suggest that the IS proposals can be improved by the information content of the UM approximate marginals, which incorporate all evidence irrespective of the observed nodes positions in the graph. However, larger betas lead to increased weight variance as in the hybrid scheme the UM marginals are not conditioned on previously sampled nodes.

| $\beta$ | 0.0 | 0.1 | 0.25 | 0.5 | 0.75 | 1.0 |
|---|---|---|---|---|---|---|
| **ESS** | 7678 | 15458 | 11779 | 1218 | 171 | 92 |