Faithful Model Inversion Substantially Improves Auto-encoding Variational Inference

Stefan D. Webb
Department of Engineering Science
Oxford University
stefan.webb@eng.ox.ac.uk

Adam Golinski
Department of Engineering Science
Oxford University
adamg@robots.ox.ac.uk

Robert Zinkov
Department of Engineering Science
Oxford University
zinkov@robots.ox.ac.uk

Frank Wood
Department of Engineering Science
Oxford University
fwood@robots.ox.ac.uk

Abstract

In learning deep generative models, the encoder for variational inference is typically formed in an ad hoc manner with a structure and parametrization analogous to the forward model. Our chief insight is that this results in coarse approximations to the posterior, and that the d-separation properties of the BN structure of the forward model should be used, in a principled way, to produce ones that are faithful to the posterior, for which we introduce the novel Compact Minimal I-map (CoMI) algorithm. Applying our method to common models reveals that standard encoder design choices lack many important edges, and through experiments we demonstrate that modelling these edges is important for optimal learning.

1 Introduction

Deep generative modelling provides models and methods for density estimation and representation learning. In a Bayesian framework, we can model the data, $\mathbf{x}$, as being generated by an unobservable, or latent, code $\mathbf{z}$, with probabilistic decoder $p_\phi(\mathbf{x} \mid \mathbf{z})$ and prior $p(\mathbf{z})$. When learning the model by stochastic gradient variational Bayes (SGVB) or the proposal distribution for inference compilation (IC), the modeller must assume a form for a probabilistic encoder $q_\psi(\mathbf{z} \mid \mathbf{x})$ approximating $p_\phi(\mathbf{z} \mid \mathbf{x})$. In other tasks, such as semi-supervised representation learning, we may begin with a probabilistic encoder, and must assume a form for the decoder and prior comprising the generative model.

Unfortunately, there exists no method for guiding the design of the structure of the desired encoder or decoder in a principled, theoretically sound way. In the deep generative modelling literature, the standard encoders are formed in an ad hoc manner by simply inverting the edges in the forward model and removing edges into the observed variables [1, 2, 3]. In the inference compilation literature, the true posterior has been represented more faithfully using a heuristic algorithm [4, 5]. For learning disentangled representations, Siddharth et al. [6] specify a structured encoder and simply use a naive Bayes model for the generative model, where the observed variables depend on all the others, the later which are fully independent from each other. These typical encoder/decoder design choices fail to encode many important conditional dependencies, and this results in suboptimal learning. Indeed, even with infinite capacity factors, an encoder that encodes independencies not in the posterior cannot learn the true posterior.

We consider models here that can be expressed as Bayesian networks (BNs), a wide class of directed Probabilistic Graphical Models (PGMs) [7] that includes latent state-space models [8, 2], sigmoid

Workshop on Advances in Approximate Bayesian Inference (NIPS 2017), Long Beach, CA, USA.

belief networks (SBNs) [9], variational autoencoders (VAEs) [1], generative adversarial networks (GANs) [10], and autoregressive networks [11, 12, 13]. PGMs is a framework for compactly representing and operating on distributions by encoding their conditional independencies, or equivalently, their factorization, in graphs, and we use it to devise a novel algorithm, CoMI, for exploiting the structure of the given model to form a compact factorization of the desired encoder or decoder that is optimal in a certain technical sense. We term such designed encoders and decoders as *minimally faithful*.

In brief, CoMI works by simulating variable elimination on the forward model in a topological ordering with a min-fill heuristic, and applying the sepset property of clique trees to determine the parents of each variable in the inverse. It has running complexity of order linear in the number of variables times the tree-width of the chosen ordering. We demonstrate CoMI on Gaussian BNs with a binary tree structure and observed leaves, for which we show how using a minimally faithful encoder results in learning a vastly closer approximation to the true posterior, holding all else constant.

2 Related work

Krishnan et al. [8] present a specific instance of our insight, that in state-space models with a single latent layer the Markov properties should be used to determine a factorization for the encoder. This implies that the encoder should use both a latent summary of the past and all subsequent observed variables as a summary of the future (or vice versa), rather than only conditioning on the previous latent variable and current observation, as in ad hoc methods. They parametrize a minimally faithful encoder using a bidirectional RNN and show that in non-linear continuous state space models superior learning is obtained, relative to using unfaithful ones.

3 Experiments

We illustrate CoMI on binary tree structures of varying depth forming Gaussian BNs—a class of models in which the cpd's are normally distributed with a fixed variance, and mean that is a fixed linear combination of its parents plus an offset. For a model of depth d,

$$X_0 \sim N(0, 1),$$
$$X_i \mid x_{\lfloor (i-1)/2 \rfloor} = y \sim N(w_i y, 1), \quad i = 1, \ldots, 2^d - 2,$$

where the $\{w_i\}$ are fixed constants sampled from $U[1/2, 2]$ and we treat the leaves $\{x_{2^{d-1}-1}, \ldots, x_{2^d-2}\}$ as the observed variables.

We performed inference compilation on trees with depth $d \in \{4, 5, \ldots, 8\}$. In the ad hoc/heuristic encoder, each inverse cpd was parametrized with a normal components and a ReLU feedforward network with $[200, 200]$ hidden units. For the minimally faithful encoder we tried initially parametrizing each factor with its own NN, but found that whilst that lead to the optimal solution being reached, quicker and more stable learning was achievable by sharing weights across the factors. For this, we developed a *novel variant* of the masked autoencoder density estimate (MADE) [11] that is able to exactly model the dependency structure of a minimally faithful encoder, and compare it against a standard MADE that uses a fully connected graph on the latent variables. We again use a ReLU feedforward network with two hidden layers for these, adjusting the number of units to match the capacity of the ad hoc/heuristic encoder.

New samples from the forward model were drawn every minibatch for training, with 25 minibatches considered to constitute an epoch, and the test objective evaluated on a single minibatch every epoch. The exact posterior under the true factorization can be calculated by using the equivalence between Gaussian BNs and multivariate normal distributions [7, §7.2]—first the forward model is converted to the parameters of a multivariate normal distribution using Theorem 7.3, which is then transformed back into a Gaussian BN for the posterior using our true factorization and Theorem 7.4. Samples from the posterior can be drawn by ancestral sampling. We evaluate inference amortization by calculating the average log-posterior of a minibatch from the encoders every epoch under five fixed datasets of the observed variables (which have not be seen by the optimizer). The learning rate was decimated when learning converged, for example, every 100 epochs in the case of $d = 5$.

The results for $d = 5$ are given in Figure 2 (other depths are similar). We observe that it is necessary to model at least the edges in an I-map for the encoder to recover the posterior (as shown by both

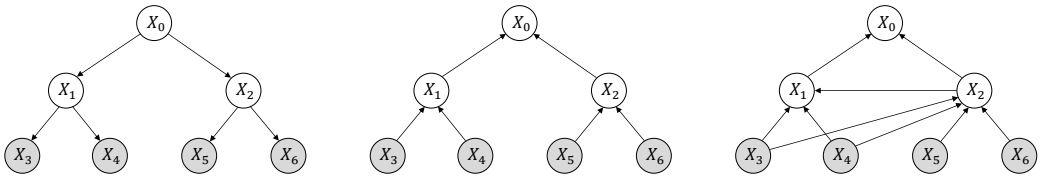

Figure 1: (a) BN structure for a binary tree with $d = 3$; (b) Inverse formed by Stuhlmuller's heuristic algorithm [4], simply inverts the edges in the generative model.; (c) Minimal I-map for the inverse produced by our algorithm, it includes an additional three edges that allow for the influence through paths via parent nodes. Therefore, the heuristic factorization is missing a third of the edges necessary to express the true inverse factorization.

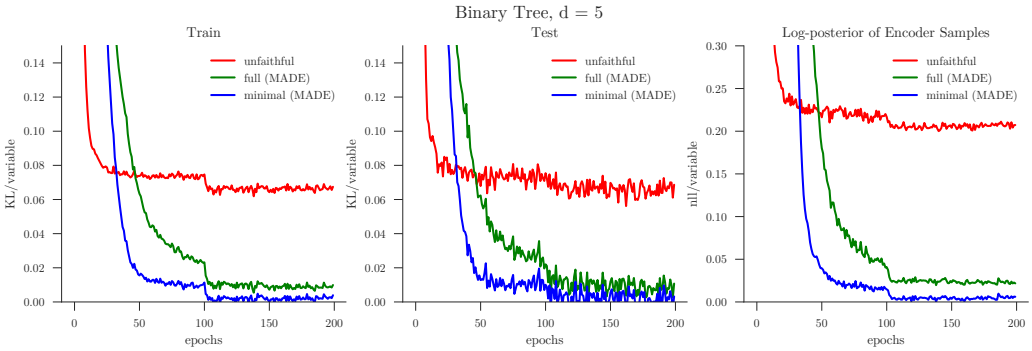

Figure 2: Comparing the reverse KL-divergence objective and log-posterior of samples from the encoders in compiled inference across factorizations and tree depths.

MADEs), and by using a minimal I-map (just the "minimal MADE") learning can be made faster and more stable. Note that we have subtracted the constant term in the IC objective, which can be calculated analytically for this model. We also observed during experimentation that if one were to decrease the capacity of all methods, learning remains stable in the minimal MADE encoder at a threshold where it becomes unstable in the other two methods.

4 Discussion

We have presented an algorithm that, given the BN structure for a generative model, produces a factorization that is a compact minimal I-map for the posterior and have argued that this should be used for the encoder. We have demonstrated that such an encoder results in learning a superior approximation to the posterior in binary trees. Future work will demonstrate the utility of this method on relaxed Bernoulli VAEs, whose ELBO is a very loose bound on the marginal log-likelihood, state-space models with multiple latent layers, and *applying the algorithm in reverse*, to produce an appropriately structured generative model given an encoder structure.

References

[1] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114v10 [stat.ML]*, 2013.

[2] Zhe Gan, Chunyuan Li, Ricardo Henao, David E Carlson, and Lawrence Carin. Deep temporal sigmoid belief networks for sequence modeling. In *Advances in Neural Information Processing Systems*, pages 2458–2466, 2015.

[3] Rajesh Ranganath, Linpeng Tang, Laurent Charlin, and David M Blei. Deep exponential families. *arXiv preprint arXiv:1411.2581v1 [stat.ML]*, 2014.

[4] Andreas Stuhlmüller, Jacob Taylor, and Noah Goodman. Learning stochastic inverses. In *Advances in neural information processing systems*, pages 3048–3056, 2013.

[5] Brooks Paige and Frank Wood. Inference networks for sequential monte carlo in graphical models. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48, 2016.

[6] N. Siddharth, Brooks Paige, Jan-Willem Van de Meent, Alban Desmaison, Frank Wood, Noah D. Goodman, Pushmeet Kohli, and Philip H. S. Torr. Learning Disentangled Representations with Semi-Supervised Deep Generative Models. 2017. URL `http://arxiv.org/abs/1706.00400`.

[7] Daphne Koller and Nir Friedman. *Probabilistic Graphical Models*. MIT Press, 2009. ISBN 9780262013192.

[8] Rahul G Krishnan, Uri Shalit, and David Sontag. Structured inference networks for nonlinear state space models. In *AAAI*, pages 2101–2109, 2017.

[9] Radford M Neal. Learning stochastic feedforward networks. *Department of Computer Science, University of Toronto*, 1990.

[10] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[11] Mathieu Germain, Karol Gregor, Iain Murray, and Hugo Larochelle. Made: masked autoencoder for distribution estimation. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pages 881–889, 2015.

[12] Aaron van den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. *arXiv preprint arXiv:1601.06759v3 [cs.CV]*, 2016.

[13] Aaron van den Oord, Nal Kalchbrenner, Lasse Espeholt, Oriol Vinyals, Alex Graves, et al. Conditional image generation with pixelcnn decoders. In *Advances in Neural Information Processing Systems*, pages 4790–4798, 2016.