Black-box Auxiliary Variational Inference

Rory Beard **Chris Lloyd** **Stephen J. Roberts**

Machine Learning Research Group
Department of Engineering Science
University of Oxford
{rory,clloyd,sjrob}@robots.ox.ac.uk

Abstract

We approach the problem of capturing the statistical dependence structure of the conditionally independent variational factors using a flexible, auxiliary variable method. We introduce two flexible auxiliary inference models; a deep neural net and a Bayesian non-paramteric model. We compare the performance of these models against each other and also consider a hybrid approach. The power of the approach lies in its ability to combine bottom-up and top-down learning in a fully Bayesian manner via a bidirectional augmented inference network.

1 INTRODUCTION

Considering at first a latent variable model with a latent variable for each datum as well as some global parameters, a traditional (mean field) variational approach is to learn a factored approximation to the posterior in which each latent variable is independently distributed and governed by its own variational parameter. Clearly this is an overly restrictive approximation since conditioning on the data statistically couples the latent variables. With careful design some statistical dependence may be retained with so-called *structured mean field* approximations (SAUL AND JORDAN, WIERGERNINCK), but the design is highly model specific and not AUTOMATICALLY SCALABLE? A natural improvement is to couple the variational factors via a common variational prior, leading to a hierarchical variational approximation (RANGANATH 2015) which are richer and more expressive. HOWEVER THE NECESSARY CRUCIAL DESIGN CHOICES MAKE THIS PROCEDURE DIFFICULT TO AUTOMATE AND SCALE?

Rather than designing model specific, parametric variational priors..., we take inspiration from the Bayesian non-parametric dimensionality reduction method of.. GP-LVM... to learn a flexible, non-linear mapping from a shared latent auxiliary space to each conditionally independent variational factor. The dependence is induced by a fixed number of latent coordinates, the locations of which are learned so as to induce the optimal dependence structure. We refer to this scheme as GP-LVM-conditional mean field (GP-LVM-CMF) inference.

From the perspective of the variational principle, we propose a drop-in replacement for mean field inference in which a rich, variational distribution is specified in the expanded space of variables we wish to infer, augmented with auxiliary inference variables. This can be seen as specifying a non-parametric hierarchical variational model (CITE RANAGNATH 2015)... integrating over the auxiliary inference parameters recovers a marginal distribution as an infinite mixture of augmented distributions...

From the perspective of recognition models and variational autoencoding, the GP-LVM-CMF scheme consists of a non-parametric inference network, or *encoder*, which maps the augmented data-auxiliary variable space to the latent variable space. ... should be more flexible than rigidly encoding directly from the non-augmented data space (WHY EXACTLY???)

In addition to this encoder model, we introduce a framework for training variational auto-encoders and their more complex derived forms. Of particular interest are deep VAE architectures that consist of stacked stochastic hidden layers in both the decoding and encoding arms. While the depth of the generative model allows for richer latent representations and generative capacity, learning the model parameters via an equivalent deep encoder leads to returns which sharply diminish as more stochastic layers are added. We demonstrate that it is the unidirectional nature of the flow of information during training that hampers the learning process, causing the hidden layers closest to the data to fit rigidly to the data, rendering model hidden units further upstream effectively inactive. To remedy this pathology we propose an auxiliary inference scheme which combines the bottom-up stochastic message passing of the vanilla encoder with top-down flowing information. We show that this Auxiliary Variational Inference (AVI) scheme signif-

icantly outperform the now standard stochastic backpropagation procedure for architectures of varying depth. The focus of this work is to inform on how to correctly learn deep, expressive generative models, rather than on the design of such generative models themselves. Therefore the experiments documented here were designed to showcase the ability of the proposed inference scheme to boost the performance of a *given* model.

Pleasingly the technique of furnishing the variational encoder with an auxiliary model is not a heuristic or hack like e.g. the popular tricks of dropout, layer-wise pre-training, gradient clipping, and more recently batch normalisation, rather it is manifest as the result of natural application of the Bayesian variational principle. In particular the technique involves further lower bounding the variational lower bound of the model evidence in a way which recovers the auxiliary lower bound considered in RERERENCE SALIMANS.

The marginal variational distribution over the variables of interest, $q(Z)$, is itself intractable precluding analyticity of the variational entropy in the lower bound therefore we lower bound again...

SVI framework permits a learning procedure that requires only unbiased estimates of the lower bound (or rather its gradient), removing the issue of the lack of analyticity of the lower bound and, for no extra cost, granting a highly data-scalable mini-batch update scheme. There has been much recent work which empirically demonstrates the success of ... despite the presence of two sources of stochasticity.

Controlling the variance of the stochastic lower bounds: reparameterization trick... theoretical Lipschitz result, ease of implementation.

We additionally introduce a kernel based approach to back-constrained amortization in which random functions are trained to capture the encoding process. WHY IS THIS GOOD: provable guarantees of kernel mappings, ameliorates pathological curvature?? This also provokes a second look at the VAE in which, in addition to replacing the decoder with a Bayesian non-parametric mapping, the same adjustment is made for the decoding part of the architecture.

We showcase the GP-LVM-CMF scheme ... Experiments:

2 BACKGROUND PRINCIPLES

2.1 VARIATIONAL INFERENCE

- Fixed form and factorised... fixed-point coordinate ascent update scheme. Loss of posterior statistical dependence - particularly bad for factor models... –¿ motivating the IBP experiments.
- Black-box approaches and SVI. Reparam trick for variance reduction, autoencoding and back-constraints for amortization.
- Augmented inference space, hierarchical variational distribution.

2.1.1 Variational Auto-encoding

- Provide a Bayesian regularisation

The VAE (2 REFERENCES) was introduced as a framework to train (potentially) deep neural net generative models in a scalable, Bayesian manner. The important insight of these authors was to reparameterise the variational variables as a deterministic transformation of a random variable drawn from a standard (parameterless) distribution so that gradient information relating to the variational parameters can be included in the stochastic gradients used to optimise the bound. Importantly this allows for variance control of the stochastic gradients, permitting practical convergence profiles, unlike the previously proposed sleep wake algorithm (REFERENCES).
—
—
— IMPORTANT VAE EQUATIONS
—
—
—

Burda et al. REFERENCE extended the VAE by including one or more stochastic hidden layers between the data and the top layer latent variables, resulting in the following hierarchical specification:

$$p_\theta(\mathbf{z}_l|\mathbf{z}_{l+1}) = \mathcal{N}(\mathbf{z}_l \mid \mu_l(\mathbf{z}_{l+1}), \sigma_l^2(\mathbf{z}_{l+1})) \quad l \in [1, L] \tag{1}$$

$$p_\theta(\mathbf{z}_L) = \mathcal{N}(\mathbf{z}_L|0, I) \tag{2}$$

where μ_l and σ_l^2 are deterministic mappings consisting of a (often 2 hidden layer) MLP. Chaining together these simple dense mappings permits a highly correlated latent dependence structure without forgoing the computational efficiency and scalability of fully factorised models.

The inference network (encoder) used in the IWAE has simply the same architecture as the generative network but with mappings in the reverse direction (figure FIGURE) i.e.

$$q_\phi(\mathbf{z}_1|\mathbf{y}) = \mathcal{N}(\mathbf{z}_1 \mid \mu(\mathbf{y}), \sigma^2(\mathbf{y})) \tag{3}$$

$$q_\phi(\mathbf{z}_l|\mathbf{z}_{l-1}) = \mathcal{N}(\mathbf{z}_l \mid \mu(\mathbf{z}_{l-1}), \sigma^2(\mathbf{z}_{l-1})) \quad l \in [2, L]. \tag{4}$$

Applying the reparameterisation trick at each stochastic layer now amounts to first sampling from a standard normal $\boldsymbol{\xi}_l \sim \mathcal{N}(0, I)$ and then applying a deterministic mapping as follows:

$$\mathbf{z}_l(\xi_l, \mathbf{z}_{l-1}, \phi) = \mu(\mathbf{z}_{l-1}, \phi) + \sigma(\mathbf{z}_{l-1}, \phi) \odot \boldsymbol{\xi}_l \tag{5}$$

Burda et al. optimise an evidence lower bound which differs from the traditional free energy in that it cannot be written as the difference between the evidence and a K-L divergence. This alternative bound highlights the fundamental similarity between variational inference and importance sampling, and in doing so begins to address the issues of the heavy penalisation under the standard bound of samples which poorly explain the data. However this penalisation worsens for layers which are further from the data as they do not receive a learning signal strong enough to overcome the K-L divergence penalisation that encourages variational posteriors towards their model priors.

2.2 AUXILIARY VARIATIONAL INFERENCE

To overcome the pathologies of the standard inference scheme, we introduce auxiliary variables $\mathbf{x}$ into the inference model as shown in figure FIGURE, which mediate a direct forward pass sampling route from the data to the top level latent variables. Since these variables do not appear in the generative model their effect must be marginalised out of the inference model. This marginalisation is however intractable and so we must further lower the bound of equation EQUATION by substituting in the Bayesian identity $q_\phi(\mathbf{z}_1, \mathbf{z}_2|\mathbf{y}) = \frac{q_\phi(\mathbf{z}_1|\mathbf{y})q_\phi(\mathbf{x}|\mathbf{y})q_\phi(\mathbf{z}_2|\mathbf{z}_1,\mathbf{x})}{q_\phi(\mathbf{x}|\mathbf{z}_1,\mathbf{z}_2,\mathbf{y})}$

2.3 THE BAYESIAN GP-LVM

The GP-LVM was proposed by Lawrence as a Bayesian non-linear dimensionality reduction model and can be viewed as a multi-output GP regression model in which the inputs are unobserved and so treated as latent variables. In the original framework these inputs were optimised rather than integrated out for tractability.

Denoting the d^{th} column of the data matrix, Y, as $\mathbf{y}_d$ the likelihood function for the GP-LVM is

$$p(Y|X) = \prod_{d=1}^{D} p(\mathbf{y}_d|X), \tag{6}$$

$$p(\mathbf{y}_d|X) = \mathcal{N}(\mathbf{y}_d|\mathbf{0}, K_{ff} + \beta^{-1}I) \tag{7}$$

where the GPs are modelled as independent across the data dimensions.

- GP-LVM
- variational compression and SVGP.
- Bayesian GP-LVM
- Provides an extremely general, Bayesian and rich way of auxiliary object upon which to condition the mean field factors in order to couple them statistically.

3 GP-LVM-CONDITIONAL MEAN FIELD VARIATIONAL INFERENCE

As is the case for the SVGP and the Bayesian GP-LVM, the inference model of the GP-LVM-CMF constructs a GP mapping to N output values which is governed by a fixed number of auxiliary inducing variables. However the motivations for this conditional inducing structure are distinct in these three cases. For the sparse GP the desire is scalability; in particular to circumvent the costly $\mathcal{O}(N^3)$ complexity of the covariance matrix inversion, which is then limited to $\mathcal{O}(N^2M)$ through variational compression. For the Bayesian GP-LVM the motivation is to be able to tractably integrate over the uncertainty associated with the unknown latent input locations. For the GP-LVM-CMF however there are no observations of the GP function values since they do not exist in the generative model, so their purpose here is to act as *pseudo-data* upon which the rest of the function values can be conditioned via the standard GP formulation.

————-THIS SHOULD POSSIBLY GO IN THE INTRO————

- Swapping out the directly parameterised $q(z)$ of a fixed-form VB distribution for the augmented GP-LVM-CMF distribution leads to an intractable lower bound.
- While it is true that we have already forsaken analyticity by performing mini-batch stochastic estimates of the lower bound - so it doesn't worsen things in the sense of analyticity that we can only sample the entropy, we still have variance control to consider.
- By judiciously further lower bounding we show that we can not only avoid such risks of large stoch gradient variance, but do so in a way which is more Bayesian than the typical case.
- This procedure results in an *auxiliary* lower bound which was inspired by independent approaches of Salimans and Welling and an UNPUBLISHED TECHNICAL NOTE on reinterpreting the sparse variational Gaussian process framework of Titsias.
- In the present work we shed new light on this auxiliary approach, and argue that it as a more correct (in the Bayesian sense) approach to inference than simply sampling the entropy term in the bound, as well as than MCMC.

We begin by applying the standard variational principle to

lower bound the marginal log likelihood as follows

$$\begin{aligned}\log p_\theta(Y) &= \mathbb{E}_{q_\phi(Z)}[\log p_\theta(Y|Z) - \log q_\phi(Z)] \\ &\quad + \mathrm{KL}[q_\phi(Z)\|p_\theta(Z|Y)] \qquad (8)\\ &\geq \mathbb{E}_{q_\phi(Z)}[\log p_\theta(Y|Z) - \log q_\phi(Z)] \\ &\triangleq \mathcal{L}_1 \qquad (9)\end{aligned}$$

where we denote the set of all fixed point generative (resp. inference) model parameters to be learned by θ (resp. ϕ), and we denote the set of all latent variables and parameters subject to inference by Z (though later we will use η to ...). Integrating 9 with respect to the auxiliary variables, noting that the marginal variational distribution over Z can be expressed as

$$q(Z) = \frac{q(Z|\mathbf{f}, U, X)q(\mathbf{f}, U, X)}{q(\mathbf{f}, U, X|Z)} \qquad (10)$$

and introducing an auxiliary model $r_\psi(\mathbf{f}, U, X)$, we arrive at the following *auxiliary lower bound*:

$$\begin{aligned}\mathcal{L}_1 &\geq \mathbb{E}_{q(Z,\mathbf{f},U,X)}\big[\log p(Y|Z)p(Z)r(\mathbf{f}, U, X|Z) \\ &\quad - \log q(Z|\mathbf{f}, X)q(\mathbf{f}, U, X)\big] \\ &\quad + \mathbb{E}_{q(Z)}\big[\mathrm{KL}[q(\mathbf{f}, U, X|Z)\|r(\mathbf{f}, U, X|Z)]\big] \qquad (11)\\ &\triangleq \mathcal{L}_{aux}. \qquad (12)\end{aligned}$$

To explain this bound, we find it useful to refer to $q(\mathbf{f}, U, X)$ (resp. $q(\mathbf{f}, U, X|Z)$) as the *variational prior* (resp.) *variational posterior* distributions over the auxiliary inference variables, the latter of which constitutes a posterior belief over the auxiliary variables having observed a (set of) sample(s), Z, generated under the measure q endowed by the inference network.

This lower bound is indistinguishable from the one we would arrive at if we were to apply the variational principle to the problem of trying to estimate (lower bound) the log evidence of the augmented model $p(Z)p(Y|Z)r(U|Z)$: getting r exactly right (i.e. choosing $r(U|Z) = q(U|Z)$) would recover $\mathcal{L}_1$ but intractability precludes knowing $q(U|Z)$ so we must replace it with a tractable surrogate, and perform inference in the augmented space rather than in the marginal Z space. –¿ Find the form of the surrogate r which leads to a variational inference task in an augmented space which is as equivalent as possible to performing variational inference in the collapsed space.

++
++
++

3.1 SCALABLE AND AMORTISED INFERENCE

Hensman SVI GVGP does do some amortising since the global inducing variables provide a statistical conduit...

Having forsaken analyticity in favour a rich and flexibly variational distribution, we must resort to gradient based methods to optimising the auxiliary lower bound with respect to the set of all parameters; generative, variational and auxiliary $\omega = \{\theta, \phi, \psi\}$, with a simulation-based stochastic approach in which we seek to optimise an MC estimate of the auxiliary lower bound:

$$\sum_{j=1}^{J}[\log p(Y|Z^{(j)})p(Z^{(j)})...] \qquad (13)$$

SVI, while introduced as a means of performing variational inference on a stochastic lower bound when the source of stochasticity is the mini-batch approach, popularised the fact that one only requires an unbiased estimate of the lower bound (in addition to satisfying some basic conditions) to converge via stochastic gradient descent on the (local) minima of the true lower bound. There has been much recent success in the way of performing so called *doubly stochastic* variational inference..... Inspired by the success of the approach in the VAE / stochastic backprop in deep networks of learning generative and inference model parameters concurrently, we perform stochastic gradient-based optimization in the joint parameter space of ω. The gradients produced by naively differentiating the MC estimate of (13) will suffer from prohibitively high variance...

variance reduction ... reparam ... A but we also note that the additional lower bounding step (11) allows the resulting auxiliary lower bound (12) to be written as

We can rearrange (12) as follows:

$$\begin{aligned}\mathcal{L}_{aux} &= \mathbb{E}_{q(Z)}[\log p(Y|Z)] \\ &\quad - \mathbb{E}_{q(\mathbf{f},U,X)}[\mathrm{KL}[q(Z|\mathbf{f}, U, X)\|p(Z)]] \\ &\quad - \mathbb{E}_{q(Z)}[\mathrm{KL}[q(\mathbf{f}, U, X)\|r(\mathbf{f}, U, X|Z)]]. \qquad (14)\end{aligned}$$

As noted by Kingma and Welling, in the case of Gaussian $p(Z)$ and $q(Z|\mathbf{f}, U, X)$, the first K-L divergence can be evaluated analytically and, but the lower lower means we can get a second analytical term... This was also noticed by Tran et al., in their lower lower bound...

Variational compression ... sparse variational GPs ... set $r(\mathbf{f}|U, X, Z) := q(\mathbf{f}|U, X)$ which means that $r(\mathbf{f}, U, X|Z) = q(\mathbf{f}|U, X)r(U, X|Z)$. Full implications of this??? Reduces configuration space... The aptness of this hinges on the veracity of the assumption that ... are sufficient statistics for ... in expectation under ...

Amortised inference ... VAE, back-constraints, ... we test several options: MLP vs kernel, q(X) vs r(U,X)...

3.2 DESIGN CHOICES

The alternative configurations...

4 RELATED WORK

Coupling mean field factors: Tran's Copula approach, the other copula approach ??, Michael Jordan paper.

The independent works of Kingma and Welling and Rezende et al. popularised the modern approach of autoencoding in the variational framework, and spurred on a great research drive in variational approaches to deep learning. Highly sophisticated models such as DRAW - a method to.... - have at their core the VAE... Grosse et al considered an alternative variational lower bound based on importance weights to train a VAE. However in all these approaches, the

deep gp - used autoencoding as backconstraints to make it scale (finite number of parameters)

Salimans and Welling considered a Markov chain inference network in which the final ($T^t h$) Markov transition in the chain produced the latent variable (sample), such that previous $T-1$ variables were treated as auxiliary variables. Collapsing over this expanded auxiliary space then resulted in a rich yet intractable mixture... and tractability was recovered through the use of an auxiliary distribution also of a Markov structure.

Most similar in spirit to GP-LVM-CMF approach is that of the *Variational Gaussian Process* developed independently by Tran et al., which was also used conditional GPs (conditioned on aux params) in the inference model to couple the factors. The key difference is that each latent variable is encoded with a *separate* conditional GP... and the coupling between the latent variables was induced by evaluating each conditional GP at the same input location. This leads to coupling whereby each GP tends to be positive in similar regions ?????. In our approach the coupling is induced via the shared inducing coordinates... more faithfully to the spirit of the GP-LVM.

5 EXPERIMENTAL SETUP

... free form q(X) means having to optimise the auxiliary latent coordinate of the test location just to be able to evaluate. Including an optimisation procedure in the predictive step is an interesting idea, but was avoided here for lack of space and clarity of discussion, so investigation was deferred to a future study.

6 DISCUSSION AND FUTURE WORK

7 TO DO:

- Graphs for the 3 models.
- Algorithm in section 3
- Discussion of the aux variables acting as a stochastic hidden layer which provides smoothing...
- Moreover, the additional KL penalisation between the variational posterior and the auxiliary distribution prevents the variational model trying to fit directly to the true posterior and instead does so via fitting its denominator (the var posterior) to the aux dist. The aux params are generated from both the latent variables and the data, acting as an intermediate between the generative likelihood model and the marginal variational model.
- Links from above to batch normalisation - the auxiliary variables act like pseudo data which are global (not batch-specific).

8 NOTES

From another perspective, the partial feedback of of z to the aux params (captured in r) means that there is a term in the cost function (aux lower bound) which is a direct effect of the generative power of aux params to generate z - an indirect is achieved when z is first passed through the generative model. So direct feedback is a reinforce step - it reinforces our current belief over z. It endows the scheme with a memory!

The above is for the aux lower lower bound in general, but we have even more motivation to use it in the case of the GP-LVM-CMF: want to avoid having to do optimization to make a prediction of a new data point!

change!!!!@

References