1st Symposium on Advances in Approximate Bayesian Inference, 2018 1–6

EP Structured Variational Autoencoders

Jonathan So JONATHAN.SO.17@UCL.AC.UK
James Townsend JAMES.TOWNSEND@CS.UCL.AC.UK
Benoit Gaujac BENOIT.GAUJAC@CS.UCL.AC.UK
University College London

Abstract

Recent work combining neural network recognition models with probabilistic graphical models has demonstrated the ability to perform fast, scalable (approximate) inference in rich, structured latent variable models that include non-linear, non-conjugate observations.

The structured variational autoencoder (SVAE) of Johnson et al. (2016) employs mean-field variational inference for computing approximate posteriors over local latents. It is often the case however, that when performing approximate inference in latent variable models, expectation propagation (EP) (Minka, 2001a,b) is able to yield more accurate posteriors than those found by mean-field, also resulting in more accurate parameters when used in a learning setting.

We present a new approach based on the SVAE framework of Johnson et al. (2016) that makes use of EP rather than mean-field for local inference. We note that an additional benefit of our approach is the greater flexibility it permits in choice of recognition network outputs. We demonstrate that this additional flexibility results in markedly improved performance over the mean-field approach on a synthetic prediction problem.

Keywords: approximate bayesian inference, expectation propagation, recognition models

1. Introduction

Probabilistic graphical models provide us with a framework for building rich structured representations in latent variable models, as well as allowing access to a number of efficient exact and approximate inference routines based on exponential family message passing (see e.g. Wainwright and Jordan, 2008 for an overview). However, in order to make use of efficient message passing routines, we are typically constrained to the class of conditionally conjugate models. For many problems of interest this class of models can prove overly restrictive.

Johnson et al. (2016) described an approach for performing efficient, scalable (approximate) inference in the class of models consisting of structured, conjugate latent variables and general non-conjugate observation likelihoods. The authors call this approach the structured variational autoencoder (SVAE). The SVAE uses mean-field variational inference (MF) in order to compute local variational parameters that optimise a surrogate objective function that is parameterised by neural network recognition models.

In this work we introduce a variant of the SVAE which utilises the expectation propagation (EP) algorithm (Minka, 2001a,b), rather than mean-field, in order to perform approximate inference over local latent variables. We refer to this approach as the EP-SVAE.

2. Background

2.1. Problem Description

We consider the class of models of the form

$$p(\theta)p(x \mid \theta)p(y \mid x, \gamma) \tag{1}$$

where θ are the global latents (parameters) corresponding to local latent variables x, and γ are the parameters of our observation likelihoods, for observed variables y. Each conditional density in (1) is individually in the exponential family, with the further restriction that $p(x \mid \theta)$ is conjugate to $p(\theta)$. Note that while the conditional density $p(y \mid x, \gamma)$ is within the exponential family, we allow for the natural parameters to depend in arbitrary non-linear ways on the sufficient statistics of $p(x \mid \theta)$, and so we lack conjugacy in the observation likelihoods.

2.2. Structured Variational Autoencoders

The SVAE targets a variational lower bound on the model evidence for the class of models considered here[1]. Similar to the stochastic variational inference (SVI) approach of Hoffman et al. (2013), the authors "optimise away" local variational parameters, which we denote τ_x, by taking them to be a function of the global variational parameters τ_θ. Unlike the SVI approach however, this function is taken to be a local partial optimiser of a *surrogate* objective function. Specifically, it returns the variational parameters found by performing mean-field variational inference on a surrogate model, in which non-conjugate observation likelihoods are replaced by conjugate recognition network potentials. The partial optimizer of this surrogate function is then denoted

$$\tau_x^*(\tau_\theta, \phi) := \arg\max_{\tau_x} \hat{\mathcal{L}}_{\text{MF}}(\tau_\theta, \tau_x, \phi) \tag{2}$$

where $\hat{\mathcal{L}}_{\text{MF}}$ is the surrogate mean-field objective and ϕ are the recognition network parameters. Note that arg max here corresponds to *any* maximum of $\hat{\mathcal{L}}_{\text{MF}}$. The use of conjugate recognition potentials permits efficient optimization of the surrogate objective by means of conjugate mean-field block co-ordinate ascent updates. The SVAE global objective then rephrases the true variational objective as a function of τ_θ, ϕ and γ,

$$\mathcal{L}_{\text{SVAE}}(\tau_\theta, \gamma, \phi) := \mathcal{L}(\tau_\theta, \tau_x^*(\tau_\theta, \phi), \gamma) \tag{3}$$

where

$$\mathcal{L}(\tau_\theta, \tau_x, \gamma) := \mathbb{E}_{q(\theta \mid \tau_\theta)q(x \mid \tau_x)}\left[\log \frac{p(\theta)p(x \mid \theta)p(y \mid x, \gamma)}{q(\theta \mid \tau_\theta)q(x \mid \tau_x)}\right] \tag{4}$$

which can be jointly optimised using auto-differentiation and stochastic gradient-based techniques.

1. Johnson et al. (2016) also allow for an exponential family prior on γ, which we omit here to ease notation.

3. Method

The SVAE chooses $\tau_x^*(\tau_\theta, \phi)$ to return a maximising argument of a surrogate mean-field objective. This appears a natural choice for the problem at hand, as the surrogate objective is similar in form to the global mean-field objective (4), the only difference being that the true observation likelihoods are replaced by conjugate recognition network potentials.

However, $\tau_x^*(\tau_\theta, \phi)$ is completely decoupled from $\mathcal{L}(\tau_\theta, \tau_x, \gamma)$ and in principle could be any function that provides a differentiable parametric mapping $(\tau_\theta, \phi) \to \tau_x$. In particular, we can choose τ_x to be the result of optimising an arbitrary surrogate objective. A desirable property of such a surrogate objective however is that it should favour solutions τ_x that result in *accurate* local posterior approximations $q(x \,|\, \tau_x)$, leading also to generative parameter estimates with low bias.

We consider here an alternative global variational objective which utilises a surrogate *EP* objective for optimising local parameters. In particular, the local variational parameters are found by iterating EP updates to convergence on a similarly defined surrogate model,

$$\hat{p}(x \,|\, \hat{\eta}_x, \phi) := p(x \,|\, \hat{\eta}_x)\psi(x; y, \phi) \tag{5}$$

where

$$\hat{\eta}_x := \mathbb{E}_{q(\theta \,|\, \tau_\theta)}\big[\eta_x(\theta)\big] \tag{6}$$

are the expected natural parameters of $p(x \,|\, \theta)$ under our current global approximation, and $\psi(x; y, \phi)$ are recognition network potentials. The global objective[2] is then defined as

$$\mathcal{L}_{\text{EPSVAE}}(\tau_\theta, \gamma, \phi) := \mathcal{L}(\tau_\theta, \hat{\tau}_x^*(\tau_\theta, \phi), \gamma) \tag{7}$$

where

$$\hat{\tau}_x^*(\tau_\theta, \phi) := \underset{\mathcal{Q}_x}{\text{ep\_fixedpoint}}[\hat{p}(x \,|\, \hat{\eta}_x, \phi)] \tag{8}$$

and $\mathcal{Q}_x$ is our approximating family. Computing $\hat{\tau}_x^*$ then corresponds to the optimisation of an EP energy function (Minka, 2005) defined by $\hat{p}$ and $\mathcal{Q}_x$.

One advantage of this formulation is that the EP-SVAE is no longer restricted to using recognition potentials $\psi(x; y, \phi)$ that are conjugate to the prior on x. EP itself provides us a mechanism for dealing with non-conjugate factors, and so in order to optimise the surrogate model efficiently, our only requirement is that we are able to easily compute the expected sufficient statistics of our approximating family under the *tilted distributions*; defined as the (normalised) product of a single factor from the target distribution $\hat{p}(x \,|\, \hat{\eta}_x, \phi)$ with the current approximation to all other factors. We demonstrate this additional flexibility in the results section.

4. Results

In order to demonstrate the EP-SVAE method we generated synthetic data according to the true generative model given below[3]

2. Note that in order to optimise (7), we require existence and differentiability of the function $\hat{\tau}_x^*(\tau_\theta, \phi)$. While this function may not be differentiable, or even defined, everywhere (EP may not converge for some τ_θ, ϕ), we proceed in the hope that encountering such situations is sufficiently rare in practice that we are able to obtain a useful algorithm nonetheless.
3. The code for these experiments is available at https://github.com/jonny-so/svae/tree/epsvae

True generative model:	Assumed generative model:
$\tau = 100$	$p(\tau) = \text{Gamma}(\tau \mid \alpha, \beta)$
$p(x_t \mid x_{t-1}, \tau) = \mathcal{N}(x_t \mid x_{t-1}, \tau^{-1})$	$p(x_t \mid x_{t-1}, \tau) = \mathcal{N}(x_t \mid x_{t-1}, \tau^{-1})$
$p(y_t \mid x_t) = \text{Bernoulli}\big(y_t \mid \Phi(2x_t)\big)$	$p(y_t \mid x_t, \gamma) = \text{Bernoulli}\big(y_t \mid \pi(x_t; \gamma)\big)$

where $t \in \{1, ..., 100\}$, Φ is the standard normal CDF, and we define $x_0 := 0$; π is a neural network with weights γ.

The true generative model above is similar to that assumed by the TrueSkill$^{\text{TM}}$ rating system of Herbrich et al. (2007) (think of a single player game, with win/loss outcome determined by the players' skill and performance alone), and we can actually perform standard EP inference efficiently in this model as this particular non-linearity allows efficient computation of the required expected statistics. However, unlike TrueSkill$^{\text{TM}}$, we make no assumption about the form of this non-linearity and instead learn it directly from data.

In the case of conjugate recognition network potentials, inference on the surrogate model for this problem requires no further approximation, and can be computed in a single forward/backward pass. In fact, in this particular case the SVAE and EP-SVAE approaches are equivalent. However, the EP-SVAE permits us the flexibility to choose alternative, non-conjugate recognition potentials.

We trained an EP-SVAE model with Mixture of Gaussian (MoG) recognition potentials using $M = 2$ mixture components and compared the prediction error on an unseen test set with that of the SVAE / conjugate potential approach. Figure 1 shows the prediction error during training for 5 independent runs of each approach. The prediction error was taken to be the negative log-probability of y_{101}, having observed $y_1, ..., y_{100}$.

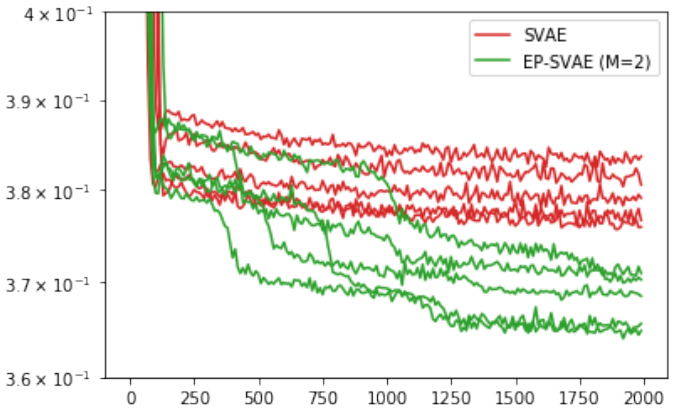


Figure 1: Log-loss prediction error (y-axis) of y_{101} having observed $y_1, ..., y_{100}$, vs number of training steps (x-axis). Displaying 5 independent runs, each averaged over 1000 sequences.

The additional flexibility provided by the MoG potentials results in a marked improvement over the prediction error achieved by the SVAE consistently over 5 independent runs. The generative likelihood for, and recognition potentials produced by an observation $y_i = 1$, as learned by the SVAE and EP-SVAE, can be seen in Appendix A.

Acknowledgments

We are grateful to Matthew Johnson for helpful correspondence and for making the SVAE code publicly available.

References

Ralf Herbrich, Thomas Minka, and Thore Graepel. Trueskill$^{\text{TM}}$: A bayesian skill rating system. In *Advances in Neural Information Processing Systems 19*, pages 569–576, 2007.

Matthew D. Hoffman, David M. Blei, Chong Wang, and John Paisley. Stochastic variational inference. *Journal Machine Learning Research (JMLR)*, 14(1):1303–1347, 2013.

Matthew J. Johnson, David Duvenaud, Alexander B. Wiltschko, Sandeep R. Datta, and Ryan P. Adams. Composing graphical models with neural networks for structured representations and fast inference. In *Advances in Neural Information Processing Systems 29*, pages 2946–2954. 2016.

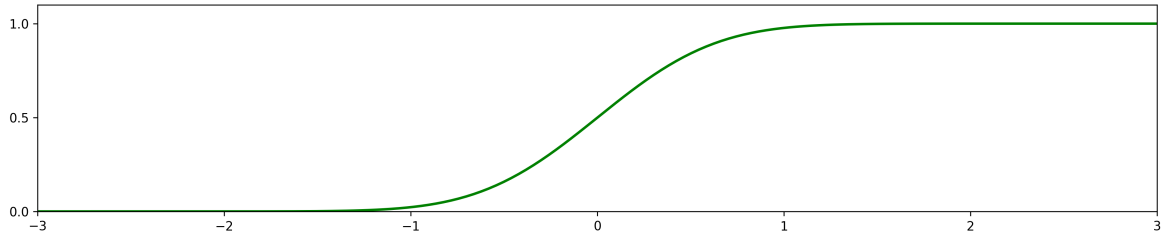
Thomas Minka. *A family of algorithms for approximate Bayesian inference*. Phd thesis, MIT, 2001a.

Thomas Minka. Expectation propagation for approximate bayesian inference. In *Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence*, pages 362–369, 2001b.

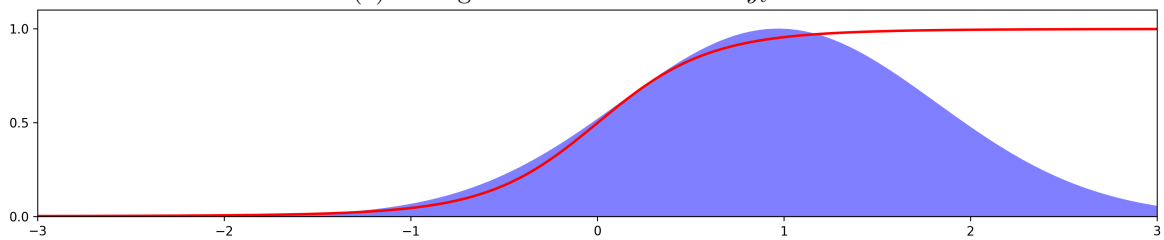
Thomas Minka. Divergence measures and message passing. Technical report, 2005.

Martin J. Wainwright and Michael I. Jordan. Graphical models, exponential families, and variational inference. *Foundational Trends in Machine Learning*, 1(1-2):1–305, 2008.

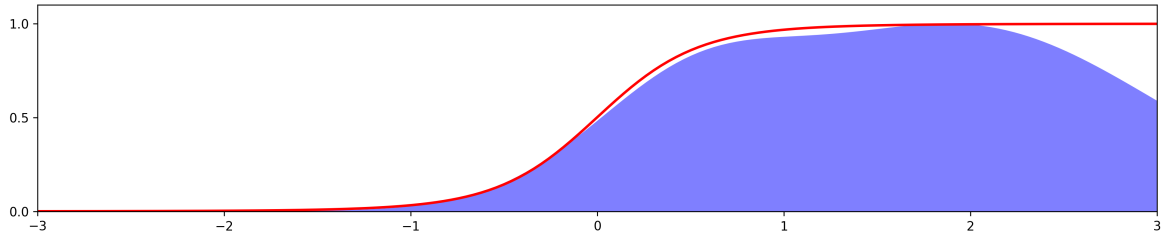
Appendix A. Observation Likelihood and Recognition Potentials



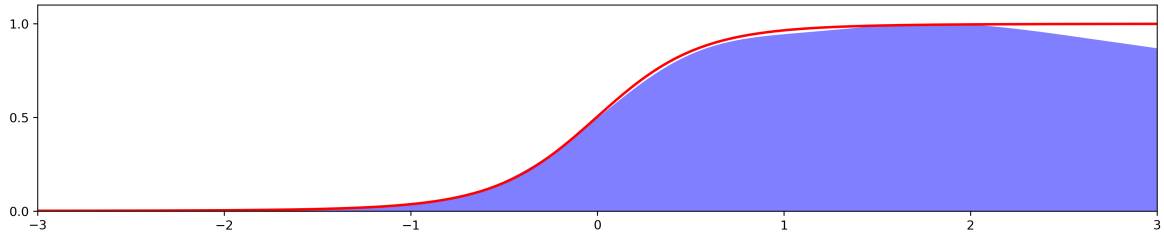
(a) True generative likelihood for $y_t = 1$



(b) SVAE



(c) EP-SVAE / MoG M=2



(d) EP-SVAE / MoG M=3

Subplot (a) shows the true generative likelihood $p(y_t = 1 \,|\, x_t)$ (green) for an observation $y_t = 1$. Subplots (b) - (d) show the equivalent learned likelihood (red) and corresponding encoder potentials (blue) for each of the trained models. All plots are displayed as a function of x_t (x-axis). Encoder potentials have been re-scaled to display a maximum value of 1.

We found that the model with $M = 3$ mixture components did not result in a significant improvement in prediction performance over $M = 2$, and so we have omitted this case from Figure 1 of Section 4 in order to keep it relatively uncluttered.