Forward Amortized Inference for Likelihood-Free Variational Marginalization

L. Ambrogioni
Radboud University

U. Güçlü
Radboud University

J. Berezutskaya
Utrecht University

E. van den Borne
Radboud University

Y. Güçlütürk
Radboud University

M. Hinne
University of Amsterdam

E. Maris
Radboud University

M. van Gerven
Radboud University

Abstract

In this paper, we introduce a new form of amortized variational inference by using the forward KL divergence in a joint-contrastive variational loss. The resulting forward amortized variational inference is a likelihood-free method as its gradient can be sampled without bias and without requiring any evaluation of either the model joint distribution or its derivatives. We prove that our new variational loss is optimized by the exact posterior marginals in the fully factorized mean-field approximation, a property that is not shared with the more conventional reverse KL inference. Furthermore, we show that forward amortized inference can be easily marginalized over large families of latent variables in order to obtain a marginalized variational posterior. We consider two examples of variational marginalization. In our first example we train a Bayesian forecaster for predicting a simplified chaotic model of atmospheric convection. In the second example we train an amortized variational approximation of a Bayesian optimal classifier by marginalizing over the model space. The result is a powerful meta-classification network that can solve arbitrary classification problems without further training.

Proceedings of the 22[nd] International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89.

1 Introduction

Bayesian inference is a principled statistical framework for estimating the probability of latent factors given a set of observations. Unfortunately, most complex Bayesian models are intractable since computing the posterior distribution involves the solution of integrals over high-dimensional spaces. Variational inference (VI) is a family of approximation methods that reframes Bayesian inference as an optimization problem that can be solved using stochastic optimization techniques [Jordan et al., 1999]. Recent developments in stochastic VI have scaled Bayesian inference to massive datasets and paved the way for the integration of deep learning and Bayesian statistics [Hoffman et al., 2013, Ranganath et al., 2014, Rezende et al., 2014, Kucukelbir et al., 2017, Tran et al., 2016]. In many applications, VI is made more efficient by optimizing a whole family of variational distributions at once [Kingma and Welling, 2013, Huszár, 2017, Ritchie et al., 2016]. This approach is usually referred to as amortized inference. Amortized inference can be seen as a special case of the larger framework of joint-contrastive variational inference [Huszár, 2017, Dumoulin et al., 2017].

In this paper we introduce forward amortized variational inference (FAVI) as a flexible and tractable new form of likelihood-free VI. FAVI is obtained by using the forward KL divergence on a joint-contrastive variational loss. One of the most important features of FAVI is that it can be used for marginalizing over a large space of nuisance variables without explicitly modeling their joint density. Marginalization of nuisance variables is important in many real-world problems such as weather forecasting [Gneiting and

Raftery, 2005]. FAVI is particularly suitable for model-based problems such as weather forecasting because it is trained on samples from the generative model. However, the applicability of FAVI goes far beyond model-based problems. As an example of a model-free problem, we use FAVI to obtain a meta-classifier as a variational approximation of the Bayes optimal classifier of an infinite ensemble of classification models. The resulting variational meta-classifier is algorithmically similar to the meta-learning methods introduced in [Prokhorov et al., 2002] and recently expanded in [Santoro et al., 2016, Vinyals et al., 2016].

2 Contributions

The main contribution of this paper is the introduction of the use of amortized variational inference with forward KL divergence and to showcase its power and flexibility in situations where a large family of nuisance variables must be marginalized out from a likelihood free model. This setting is of great importance in fields such as physics and meteorology where it is not usually possible to backpropagate through the generative models and where there are a large number of nuisance variables. While the stochastic loss used in this paper has already been used as a sub-component of importance sampling schemes [Le et al., 2017a, Papamakarios and Murray, 2016], to the best of our knowledge we are the first to contextualize this method in the framework of variational inference. Furthermore, we introduce two new algorithms based on the idea of likelihood free marginalization. The first algorithm is a Bayesian forecaster that can be used with arbitrary signal and noise models. The second, more ambitious, algorithm is a Bayesian meta-learning system obtained as a variational approximation of the Bayes optimal classifier of an ensemble. We also show that our method outperforms conventional variational inference even in settings where an analytic likelihood is available.

3 Related work

In spite of its theoretical advantages, the intractability of the expectation in the forward KL divergence $D_{KL}(p(z \mid x) \| q(z))$ limits its applicability in the conventional VI framework [Bishop, 2006]. The forward KL is adopted by expectation propagation (EP) methods [Minka, 2001, Barthelmé and Chopin, 2011], but EP is not a form of VI since it does not minimize a global divergence between the two distributions. Likelihood-free Bayesian inference is often based on approximate Bayesian computation (ABC) [Tavaré et al., 1997, Pritchard et al., 1999]. Recently the ABC approach has been applied to both VI [Tran et al., 2017a] and EP [Barthelmé and Chopin, 2011]. However, despite its success in many applications, ABC has some important limitations. In particular, the efficiency of rejection-based ABC methods tends to sharply degrade as the dimensionality grows and the use of low-dimensional summary statistics can severely affect the performance. Similarly, methods based on some form of density estimation such as [S. et al., 2018] are strongly affected by the curse of dimensionality since high-dimensional density estimation is notoriously challenging. An alternative approach, which is algorithmically similar to our method, is to treat Bayesian inference as a nonlinear regression problem. This approach was first introduced in [Blum and François, 2010] and recently extended in [Papamakarios and Murray, 2016] and [Le et al., 2017a]. In this latter work, a loss similar to our FAVI loss was iteratively optimized using an importance sampling scheme so that the simulator ($p(z, x)$ in our notation) gradually narrows down to the distribution of the observed data. Note that this work does not draw any connection with VI and their importance sampling scheme is explicitly designed to avoid inference amortization. In general, the FAVI approach offers a theoretical foundation to several previous works based on training deep networks on simulated data [Le et al., 2017b, Jaderberg et al., 2014, 2016, Gupta et al., 2016, Stark et al., 2015, Güçlütürk et al., 2016, Ambrogioni et al., 2017a].

Most of the recent literature about likelihood-free approximate Bayesian inference is based on adversarial training. This line of research was initiated by adversarially learned inference (ALI) which can be shown to minimize the Jensen-Shannon divergence at the limit of an optimal discriminator [Dumoulin et al., 2017]. Several other adversarial VI methods have recently been introduced [Mescheder et al., 2017, Tran et al., 2017b, Huszár, 2017]. These variational methods share some of the flexibility of FAVI, but they usually require the samples from p to be differentiable. A drawback of adversarial methods is that the adversarial minimax problem is equivalent to the minimization of a divergence only in the nonparametric limit [Goodfellow et al., 2014, Mescheder et al., 2017]. From a practical perspective, variational methods tend to generate very realistic samples, but often suffer from instability during training and mode collapse [Arora et al., 2018, 2017].

4 Background on joint-contrastive variational inference

Joint-contrastive variational inference was first introduced in the context of ALI [Dumoulin et al., 2017] and more explicitly outlined in [Huszár, 2017]. The loss

functional of joint-contrastive variational inference is a divergence between the model joint distribution and a joint variational distribution:

$$\mathcal{L}_j[p,q] = D(p(z,x)\|q(z,x)) \ . \tag{1}$$

Without further constraints the minimization of this loss functional is not particularly useful as the model joint $p(z,x)$ is usually tractable and it does not need to be approximated. The key idea for approximating the intractable posterior $p(z \mid x)$ by minimizing Eq. 1 is to factorize the variational joint as the product of a variational posterior $q(z \mid x)$ and the sampling distribution of the data:

$$q(x,z) = q(z \mid x)k(x) \ . \tag{2}$$

Usually $k(x)$ is a re-sampling distribution of a training set as in the case of variational autoencoders [Kingma and Welling, 2013]. Given this factorization, the minimization of Eq. 1 with respect to both q and p simultaneously approximates the model posterior with $q(z \mid x)$ and the real-word distribution with $p(x)$. Importantly, we can often sample from both $q(x,z)$ and $p(z,x)$ and this implies that we can stochastically optimize Eq. 1 for a large class of divergence measures.

4.1 Amortized inference

If we adopt the KL divergence in Eq. 2, the joint-contrastive variational inference loss decomposes into an evidence loss and an amortized inference loss term:

$$\begin{aligned} &D_{KL}(q(x,z)\|p(x,z)) = \\ &D_{KL}(k(x)\|p(x)) + \mathbb{E}_{k(x)}[D_{KL}(q(z \mid x)\|p(z \mid x))] \ . \end{aligned} \tag{3}$$

The result suggests that conventional amortized inference is a special case of joint-contrastive variational inference. We can see this by studying the gradients of Eq. 3. In the following, ∇_q denotes the functional gradient with respect to the density q. We use this functional notation in order to avoid referring to an explicit parametrization. Since the term corresponding to the entropy of $k(x)$ in Eq. 3 does not depend on q, this divergence has the same functional gradient as the (negative) amortized ELBO:

$$\begin{aligned} &\nabla_q D_{KL}(q\|p) \\ &= \nabla_q \mathbb{E}_{q(x,z)}\left[\log \frac{q(z \mid x)}{p(x,z)}\right] + \nabla_q \mathbb{E}_{k(x)}[\log k(x)] \\ &= -\nabla_q \mathbb{E}_{k(x)}[\text{ELBO}(q,p)] \ . \end{aligned} \tag{4}$$

Therefore, amortized variational inference is a special case of joint-contrastive variational inference.

5 Forward amortized variational inference

The reverse KL divergence has a central position in the classical (posterior-contrastive) variational framework because it leads to a tractable variational lower bound. Conversely, the forward KL divergence is intractable in a posterior-contrastive sense as it requires computation of an expectation with respect to the true posterior. We will now show that the forward KL is tractable when used in a joint-contrastive loss. In this case we obtain the following divergence:

$$\begin{aligned} &D_{KL}(p(x,z)\|q(x,z)) \\ &= \mathbb{E}_{p(x,z)}\left[\log \frac{p(x,z)}{q(z \mid x)k(x)}\right] \\ &= -\mathbb{E}_{p(x,z)}[\log q(z \mid x)] + \mathbb{E}_{p(x,z)}\left[\log \frac{p(x,z)}{k(x)}\right] \ . \end{aligned} \tag{5}$$

Note that in this expression there is only one term that depends on q. Therefore, by ignoring the constant terms, we can define the FAVI loss as follows:

$$\mathcal{L}_{FA} = -\mathbb{E}_{p(x,z)}[\log q(z \mid x)] \ . \tag{6}$$

The resulting functional gradient is given by

$$\nabla_q \mathcal{L}_{FA} = -\mathbb{E}_{p(x,z)}[\nabla_q \log q(z \mid x)] \ . \tag{7}$$

Note that the computation of this gradient requires neither reparametrization tricks nor black-box methods, since the expectation is taken with respect to p while the gradient is taken with respect to q.

The FAVI variational loss can also be derived as an amortized form of posterior-contrastive variational inference. The forward KL posterior-contrastive variational loss is given by:

$$D_{KL}(p(z \mid x)\|q(z \mid x)) = \mathbb{E}_{p(z|x)}\left[\log \frac{p(z \mid x)}{q(z \mid x)}\right] \ . \tag{8}$$

It is challenging to obtain unbiased samples from the gradient of this expression as the expectation is taken with respect to the intractable $p(z \mid x)$. We can recover the FAVI loss (up to a term constant in q) if we amortize the problem with respect to the model probability:

$$\begin{aligned} &\mathbb{E}_{p(x)}[D_{KL}(p(z \mid x)\|q(z \mid x))] \\ &\quad = -\mathbb{E}_{p(x,z)}[\log q(z \mid x)] + \mathbb{E}_{p(x,z)}[\log p(x,z)] \ . \end{aligned} \tag{9}$$

FAVI has several advantages over reverse amortized inference. First of all, it is very simple to obtain Monte Carlo samples of the gradients of the stochastic loss in Eq. 6, since the expectation is taken with respect to p. This avoids the use of methods such

as the reparametrization trick, which limits the family of possible probability distributions and lengthens the computational graph since the loss needs to be back-propagated through the samples [Rezende et al., 2014]. Another important advantage is that the model joint probability $p(x, z)$ does not need to be evaluated explicitly. This implies that FAVI can be used when the likelihood is intractable, in situations where ABC methods are usually adopted [Csilléry et al., 2010, Blum and François, 2010, Marin et al., 2012, Tran et al., 2017a]. A downside of FAVI is that Eq. 5 cannot be directly minimized with respect to p since $k(x)$ cannot be expressed in closed form. There are several possible ways for dealing with this problem. In Appendix A we outline an adversarial method that only requires the differentiability of the samples from p. Note that the optimization of p is not strictly speaking part of Bayesian inference. Therefore we will focus on the case where the generative model p is known *a priori* in the rest of the paper. One of the most interesting features of FAVI is that its loss is optimized by the exact marginals even when the variational approximation is fully factorized, as we shall demonstrate in the next section.

5.1 Marginalization properties of FAVI

In the fully factorized mean field approximation the FAVI loss is minimized by the exact marginals of the true posterior, as stated in the following theorem:

Theorem 1 (Exact marginals)**.** *Consider a joint distribution $p(z, x)$ and a fully factorized variational posterior $q(z \mid x) = \prod_k q_k(z_k \mid x)$. The functional $\mathcal{L}_{FA}[p, q]$ is minimized when $q(z \mid x) = \prod_k p(z_k \mid x)$ for all x in the support of $p(x)$. Furthermore, the minimizer is unique when all values of x are in the support of $p(x)$.*

Proof. In the fully factorized case, the FAVI loss can be rewritten as follows:

$$\begin{aligned}\mathcal{L}_{FA} &= -\mathbb{E}_{p(x,z)}\left[\sum_k \log q_k(z_k \mid x)\right] \\ &= -\sum_k \mathbb{E}_{p(z_k|x)p(x)}[\log q_k(z_k \mid x)] \\ &= \sum_k \mathbb{E}_{p(x)}[D_{KL}(p(z_k \mid x) \| q_k(z_k \mid x))] \\ &\quad - \sum_k \mathbb{E}_{p(z_k|x)}[\log\left(p(z_k \mid x)\right)] \ . \end{aligned} \tag{10}$$

The conditional entropy term on the right side of the final expression does not depend on q and can therefore be ignored. Since the KL divergence is always non-negative and vanishes only when the two distributions are identically equal, the expectations in the remaining term are equal to zero if and only if $q_k(z_k \mid x) = p(z_k \mid x)$ for all k and for all x in the support of $p(x)$. □

The situation is radically different in reverse KL VI where the factorized approximation can lead to a severe underestimation of the uncertainty of the marginals [Murphy, 2012, Bishop, 2006].

Theorem 1 straightforwardly generalizes to variational models that are factorized into two blocks. From this, an important result follows:

Theorem 2 (Consistent marginalization)**.** *Consider a joint distribution $p(z, \xi, x)$ and the (nonparametric) conditionally independent variational model $q(z, \xi \mid x) = q_z(z \mid x) q_\xi(\xi \mid x)$. The following equality holds:*

$$\begin{aligned}&\int \operatorname*{arginf}_q \mathcal{L}_{FA}[p(z, \xi, x), q(z, \xi \mid x)] d\xi \\ &= \operatorname*{arginf}_{q_z} \mathcal{L}_{FA}[p(z, x), q_z(z \mid x)] \ . \end{aligned} \tag{11}$$

Proof.

$$\begin{aligned}&\int \operatorname*{arginf}_q \mathcal{L}_{FA}[p(z, \xi, x), q(z, \xi \mid x)] \, \mathrm{d}\xi \\ &= \int p(z \mid x) p(\xi \mid x) \, \mathrm{d}\xi \\ &= p(z \mid x) = \operatorname*{arginf}_{q_z} \mathcal{L}_{FA}[p(z, x), q_z(z \mid x)] \ , \end{aligned} \tag{12}$$

where the first equality is a direct consequence of Theorem 1. □

Therefore, there is no need to explicitly model the conditional dependencies between z and ξ when the aim is to estimate $q_z(z \mid x)$. In practice, it is straightforward to obtain Monte Carlo estimates of the marginalized variational loss $\mathcal{L}_{FA}[p(z, x), q_z(z \mid x)]$, since a sample from $p(z, x)$ is obtained by 'ignoring' ξ from a sample from the full joint distribution. Conversely, marginalization in the reverse KL approach requires to either perform the challenging integration $p(z, x) = \int p(z, \xi, x) d\xi$ or to explicitly model the conditional dependencies between z and ξ and marginalize out ξ from the resulting variational distribution. Note that Theorem 2 does not hold in the case of reverse KL inference and, consequently, assuming conditional independence could severely bias the resulting marginalized posterior.

6 Applications

We begin this section with a direct comparison between amortized reverse VI and FAVI. In this comparison we approximate the variational posterior of

a variational autoencoder and we compare the accuracy of the two variational posteriors. Subsequently, we discuss two applications where the reverse KL approach is not easily applicable. These applications involve large-scale likelihood-free marginalization of latent variables. In the first application we use FAVI to obtain a variational forecaster of chaotic time series. This is an example of a model-based problem since the dynamic equations are assumed to reliably describe the dynamics of real-world systems such as the earth's atmosphere. In the second application we apply FAVI to the model-free problem of meta-classification. In this case, predictive performance is obtained by marginalizing over the posterior distribution of a weakly structured ensemble of random classification models that span a very large space of possible classification problems.

6.1 Comparison between amortized inference methods

In order to compare FAVI with other amortized inference approaches, we approximated the posterior distribution $p(\boldsymbol{z} \mid \boldsymbol{x})$ of the generative model

$$p(\boldsymbol{x}, \boldsymbol{z}) = \mathcal{N}\big(\boldsymbol{x}; \boldsymbol{f}_\mu(\boldsymbol{z}), \mathrm{diag}(\boldsymbol{f}_\sigma(\boldsymbol{z}))\big)\mathcal{N}\big(\boldsymbol{z}; \boldsymbol{0}, I\big) \ ,$$

where $\boldsymbol{x}$ is a vector containing the intensity of the pixels of a black-and-white image and $\boldsymbol{z}$ is a vector of latent variables. The functions $\boldsymbol{f}_\mu(\boldsymbol{z})$ and $\boldsymbol{f}_\sigma(\boldsymbol{z})$ are the two outputs of a pre-trained deep neural network. The network has a three-layered fully connected architecture with ReLu nonlinearities in the hidden layers and was trained on the MNIST dataset using a variational autoencoder [Kingma and Welling, 2013]. We decided to use a common pre-trained generator in order to have a clean comparison between the performances of the approximate Bayesian inference methods. The variational posterior $q(\boldsymbol{z} \mid \boldsymbol{x})$ was parametrized by a three-layered fully connected architecture with ReLu nonlinearities. Both models trained with Adam [Kingma and Ba, 2014] for 100 epochs with batch size 200. The reverse KL inference network was trained by re-sampling MNIST images while FAVI was trained on simulated samples. This difference follows from the fact that the former is amortized with respect to $k(x)$ while the latter is amortized with respect to $p(x)$. We also included ALI [Dumoulin et al., 2017] in this comparison as an example of an adversarial likelihood-free method.

6.1.1 Results

The latent reconstruction error was quantified as respectively

$$\mathbb{E}_{q(\hat{\boldsymbol{z}}|\boldsymbol{x})p(\boldsymbol{x},\boldsymbol{z})}\left[\frac{1}{N}\sum_{j=1}^{N}(z_j - \hat{z}_j)^2\right]$$

and

$$\mathbb{E}_{p(\hat{\boldsymbol{x}}|\hat{\boldsymbol{z}})q(\hat{\boldsymbol{z}}|\boldsymbol{x})k(\boldsymbol{x})}\left[\frac{1}{M}\sum_{j=1}^{M}(x_j - \hat{x}_j)^2\right] \ ,$$

where N is the dimension of the latent space and M is the number of pixels. We tested the statistical difference between the errors using two-sample t-tests. Figure 1A shows the reconstruction error of the latent variable given a generated image. As we can see, FAVI has a remarkably lower latent reconstruction error when compared with reverse KL VI ($p < 0.001$). The latent error of ALI is slightly smaller than the error of reverse KL VI ($p < 0.001$). The superior performance of FAVI could have been expected since FAVI is trained on generated images while the reverse KL method is trained directly on real data. However, FAVI also has a slightly lower and less variable observable reconstruction error ($p < 0.05$). This can be seen in Fig. 1B. Conversely, ALI has a very high reconstruction error. Figure 1B shows the two variational distributions of the first component of the latent vector for an example image.

6.2 Bayesian variational forecaster

Forecasting the future of a dynamical system based on past noisy measurements and a system of dynamic equations is crucial for many scientific applications [West, 1996]. The most well-known of these applications is arguably weather forecasting [Gneiting and Raftery, 2005]. FAVI is particularly appropriate for dynamic forecasting problems for three main reasons. First, in these problems the generator is known with good accuracy and this benefits approaches like FAVI where the training samples are sampled from the generator. Second, it is often difficult to obtain analytic expressions for the probability densities of the dynamic and the noise models. Third, forecasting highly benefits from the marginalization of nuisance variables and unknown parameters [Gneiting and Raftery, 2005]. We validated our FAVI forecaster on a simulated dataset. We generated chaotic time series using a very simplified model of atmospheric convection: the Lorentz dynamical system [Lorenz, 1963]. The system is given by the following differential equations:

$$\begin{aligned}
\dot{x}_1(t) &= 10\ \big(x_2(t) - x_1(t)\big) \\
\dot{x}_2(t) &= x_1(t)\big(28 - x_3(t)\big) - x_2(t) \\
\dot{x}_3(t) &= x_1(t)x_2(t) - 8/3x_3(t) \ ,
\end{aligned}$$

where the dot denotes a derivative with respect to time. In our case, the task is to estimate the probability of the value of x_1 at the future time point t^* given a set of M noise-corrupted observations $Y =$

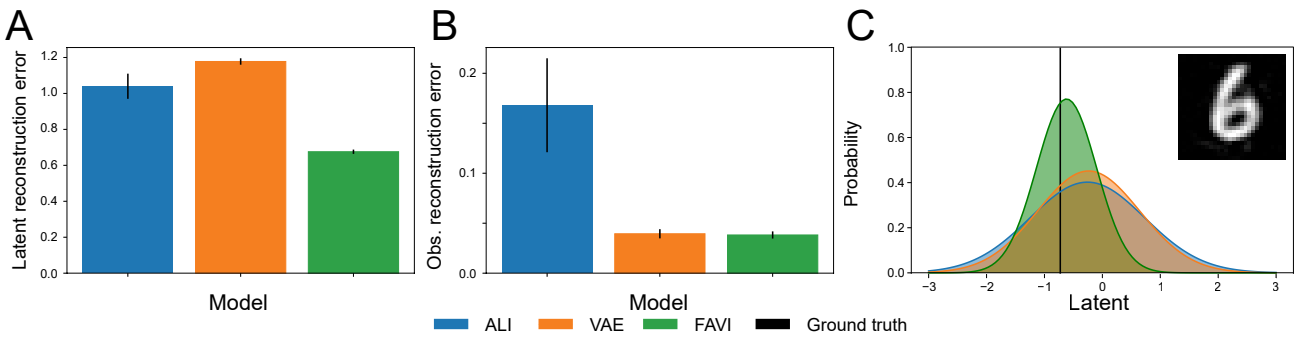

Figure 1: Comparison of the performance of the variational inference methods on the MNIST dataset. **A.** Reconstruction error of the latent variables. **B.** Reconstruction error of the images. **C.** Example of variational distributions given the (synthetic) image shown in the upper right corner. The black line denotes the real value of the latent variable.

$\{y(t_0), ..., y(t_M)\}$ where

$$y(t) \sim \mathcal{N}(x_1(t), 10^2) \ . \tag{13}$$

Note that the variables x_2 and x_3 are not observed and need to be marginalized out. The graphical model of the complete and marginalized joint is given in Fig. 2A. The FAVI loss is given by:

$$\mathcal{L}_{FA} = -\mathbb{E}_{p(x(t^*),Y)}[\log q\left(x(t^*) \mid Y\right)] \ . \tag{14}$$

We parametrized $q\left(x(t^*) \mid Y\right)$ using a dilated convolutional neural network [Yu and Koltun, 2015] with a kernel mixture network output [Ambrogioni et al., 2017b], the details of the architecture are given in Appendix B.

6.2.1 Results

We compared the performance of our variational Bayesian forecaster against the extended Kalman filter (EKF), one of the most popular off-the-shelf dynamic forecasting methods [Evensen, 2009]. Specifically, we used the EKF for obtaining the joint posterior probability density of each variable at the last time point t_M given the observations. By construction of the EKF approximation, this probability is a multivariate normal distribution. We made a forecast by numerically integrating 500 time series from t_M to t^*, where the initial conditions were sampled from the EKF posterior density at t_M. Figure 2B shows the forecast of a randomly sampled example trial together with the ground truth. The predictive distribution of the Bayesian variational forecaster is tightly tracking the ground truth. Interestingly, the variational posterior bifurcates into the two possible 'wings' of the Lorentz attractor. For each validation trial the performances of the EKF and the variational Bayesian forecaster were quantified as the probability of $x_1(t^*)$ being inside a symmetric interval centered around the ground truth with radius 3. In the EKF case this probability was obtained by counting the number of samples inside the interval and dividing by the total number of samples, while in the case of the Bayesian variational forecaster the probability was obtained by integrating the variational posterior probability density inside the interval. Figure 2C shows the scatter plot of these probabilities for 500 validation trials. On average the performance of the variational Bayesian forecaster is 1.94 times higher than the performance of the EKF.

6.3 Bayesian variational meta-classifier

We now introduce a real-wold application that showcases the flexibility and scalability of FAVI when the real generative model is unknown. Our aim is to construct a Bayesian meta-classifier as an amortized variational approximation of the Bayes optimal classifier of an ensemble. Conventional variational methods are not suited for this task as they would need to introduce a variational distribution over the potentially infinite and unstructured model space and explicitly marginalize over the resulting posterior. Furthermore, the model likelihood $p(D \mid M_k)$ is very often non-differentiable and even impossible to evaluate in closed form. The lack of differentiability would rule out adversarial variational methods. We begin by giving a brief introduction to ensemble methods and Bayes optimal classifiers.

6.3.1 Bayesian ensembles

In a classification task the aim is to estimate the probability of the target class assignments y given a set of predictors $\boldsymbol{x}$. In an ensemble learning setting we assume that the classification task is sampled from a pre-

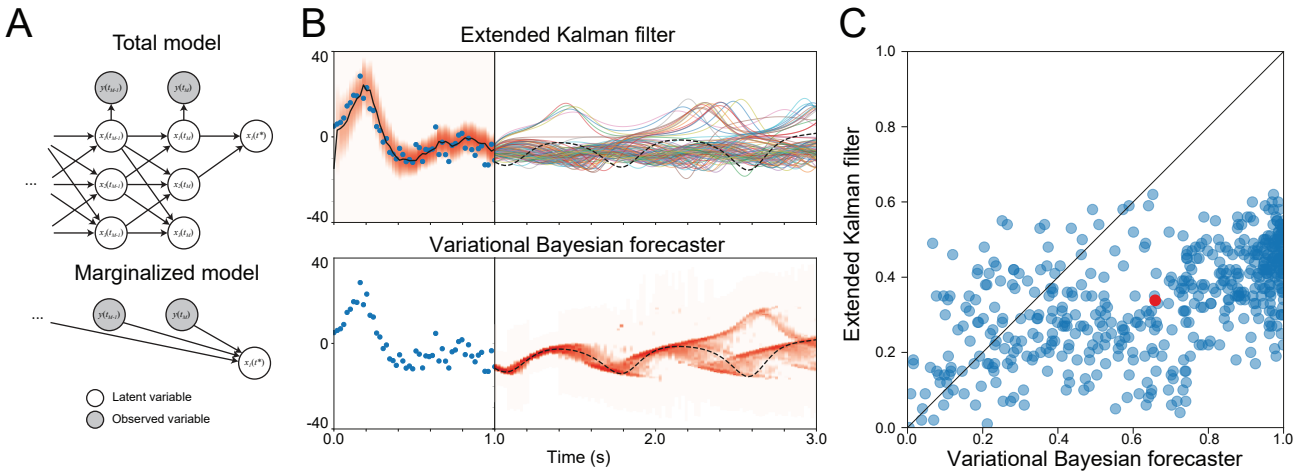

Figure 2: **A.** The total and marginalized generative models for forward autoencoders. **B.** EKF (top panel) and variational (bottom panel) forecast of a time series sampled from the Lorentz system. The blue dots denote the noise-corrupted observations. **C.** Forecast of a Lorentz dynamical system. The blue dots are individual simulated trials and the red dot denote the mean (center of mass).

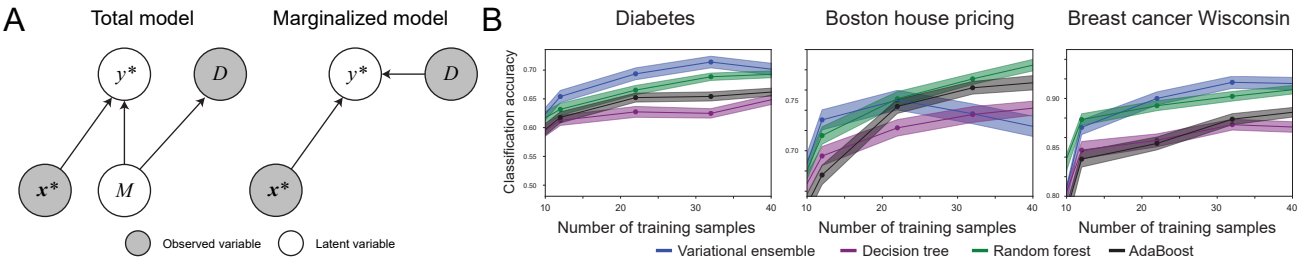

Figure 3: **A.** The total and marginalized generative models for the Bayesian meta-classifier. **B.** The classification accuracy of the Bayes variational meta-classifier versus three common alternatives on three data sets.

defined family of classification models $M_1, M_2, ..., M_K$. In our notation we consider two models with the same parametric form, but different parameter values as different models. An ensemble classifier has the following form:

$$p(y^* \mid \boldsymbol{x}^*, D) = \sum_{k=1}^{K} w_k(D) p(y^* \mid \boldsymbol{x}^*, M_k) \ , \qquad (15)$$

where x^* denotes a new vector of predictors, y^* denotes the corresponding label and D denotes the training data. Different ensemble models use different techniques for setting the weights $w_k(D)$. The optimal way of setting the weights $w_k(D)$ can be obtained formally using Bayes' rule. The posterior probability of each model M_k given the training data is given by:

$$p(M_k \mid D) = \frac{p(D \mid M_k) p(M_k)}{p(D)} \ , \qquad (16)$$

where D is a training set of predictors $\boldsymbol{x}$ and target class assignments y. Assuming that we know the prior over the family of classification models, the optimal solution to the classification problem is given by marginalizing the posterior distribution $p(y \mid \boldsymbol{x})$ over all models $M_1, M_2, ..., M_K$ [Mitchell, 1997]. This is known as the Bayesian optimal classifier:

$$p(y^* \mid \boldsymbol{x}^*, D) = \sum_{k=1}^{K} p(y^* \mid \boldsymbol{x}^*, M_k) p(M_k \mid D) \ . \qquad (17)$$

In practice, computing the Bayesian optimal classifier is intractable as it involves a sum (or an integral) over the whole (usually infinite) ensemble of models.

6.3.2 Variational meta-classifier

A Bayesian variational meta-classifier can be obtained by approximating the Bayesian optimal classifier using FAVI. The model is amortized with respect to whole training sets consisting of feature/label pairs, which are assumed to be generated by one (and only one) of the models in the ensemble. The resulting amortized posterior model is a meta-classifier, as it takes as input a training set and it outputs the predictive distribution over the label of an arbitrary new data-point. The

forward amortized loss is given by:

$$\mathcal{L}_{FA}[q] = -\mathbb{E}_{p(y^*,x^*,D)}[q(y^* \mid x^*, D)] \ ,$$

where

$$p(y^*, x^*, D) = \sum_k p(y^*, x^*, D \mid M_k)p(M_k) \ .$$

The graphical models of both the total and the marginalized joint are shown in Fig. 3A. We trained a recurrent neural network (RNN) using the FAVI loss in order to approximate the predictive distribution $p(x \mid y)$. Our variational posterior is given by:

$$q(y^* = 1 \mid \boldsymbol{x}^*) = \text{RNN}(\boldsymbol{x}^*; D) \ , \tag{18}$$

where $\text{RNN}(\boldsymbol{x}^*; D)$ is a recurrent architecture that has received as input a training set D of training pairs $(\boldsymbol{x}, y)$. The details of our RNN architecture are given in Appendix C.

6.3.3 Results

We trained a Bayes variational meta-classifier model on the ensemble of generative models described in Appendix D. The network was trained on binary classification with 10 predictors. The Chainer deep learning framework [Tokui et al., 2015] was used for model training. After training the model was tested separately on three public real-world datasets: the Boston house-prices dataset, the diabetes dataset and the breast cancer Wisconsin dataset [Harrison and Rubinfeld, 1978, Efron et al., 2004, Street et al., 1993]. In all datasets only the first 10 predictors were used. The Boston dataset is a regression problem but we converted it into a classification problem by replacing the value of the output variable with label 0 if it was less than the total median or with label 1 otherwise. The datasets contained 506, 442 and 569 data points, respectively. However, in order to reliably evaluate the model performance on small data, in each dataset we sampled data subsets of length N (from $N = 12$ to $N = 42$) at random. The model was tested by making a prediction for the $(N + 1)$-th sample. The sampling and testing was repeated 500 times for different re-samplings of the full dataset and the model performance scores were averaged. The model performance was compared to three other models: random forest, AdaBoost and decision trees [Freund et al., 1999, Breiman, 2001, Quinlan, 1986, Dietterich, 2000]. Our experiments show that the Bayesian variational meta-classifier is competitive when compared to other ensemble approaches, achieving the best performances in diabetes and breast cancer datasets (Fig. 3B). In the house pricing dataset the Bayesian variational meta-classifier has competitive performance when the training set is smaller than 20 data-points, but the performance degrades for higher number of training samples. This decline in performance is likely to be caused by the limitations of our recurrent architecture. Note that the variational meta-classifier is applied to each dataset without any further training, while the other methods are trained separately on each dataset.

7 Conclusions

In this paper we introduced a likelihood-free variational method based on the minimization of the forward KL divergence between the model joint distribution and a factorized variational joint distribution. We focused our exposition on variational marginalization problems where a Bayesian predictive distribution is obtained by marginalizing over a large space of latent variables.

References

M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37(2): 183–233, 1999.

M. D. Hoffman, D. M. Blei, C. Wang, and J. Paisley. Stochastic variational inference. *The Journal of Machine Learning Research*, 14(1):1303–1347, 2013.

R. Ranganath, S. Gerrish, and D. Blei. Black box variational inference. *International Conference on Artificial Intelligence and Statistics*, 2014.

D. J. Rezende, S. Mohamed, and D. Wierstra. Stochastic backpropagation and approximate inference in deep generative models. *International Conference on Machine Learning*, 2014.

A. Kucukelbir, D. Tran, R. Ranganath, A. Gelman, and D. M. Blei. Automatic differentiation variational inference. *The Journal of Machine Learning Research*, 18(1):430–474, 2017.

D. Tran, A. Kucukelbir, A. B. Dieng, M. Rudolph, D. Liang, and D. M. Blei. Edward: A library for probabilistic modeling, inference, and criticism. *arXiv preprint arXiv:1610.09787*, 2016.

D. P. Kingma and M. Welling. Auto-encoding variational Bayes. *arXiv preprint arXiv:1312.6114*, 2013.

F. Huszár. Variational inference using implicit distributions. *arXiv preprint arXiv:1702.08235*, 2017.

D. Ritchie, P. Horsfall, and N. D. Goodman. Deep amortized inference for probabilistic programs. *arXiv preprint arXiv:1610.05735*, 2016.

V. Dumoulin, I. Belghazi, B. Poole, O. Mastropietro, A. Lamb, M. Arjovsky, and A. Courville. Adversarially learned inference. *International Conference on Learning Representations*, 2017.

T. Gneiting and A. E. Raftery. Weather forecasting with ensemble methods. *Science*, 310(5746):248–249, 2005.

D. V. Prokhorov, L. A. Feldkarnp, and I. Y. Tyukin. Adaptive behavior with fixed weights in RNN: an overview. *International Joint Conference on Neural Networks*, 3, 2002.

A. Santoro, S. Bartunov, M. Botvinick, D. Wierstra, and T. Lillicrap. Meta–learning with memory–augmented neural networks. *International Conference on Machine Learning*, 2016.

O. Vinyals, C. Blundell, T. Lillicrap, and D. Wierstra. Matching networks for one shot learning. *Advances in Neural Information Processing Systems*, 2016.

T. A. Le, A. G. Baydin, and F. Wood. Inference compilation and universal probabilistic programming. *AISTATS*, 2017a.

G. Papamakarios and I. Murray. Fast epsilon-free inference of simulation models with Bayesian conditional density estimation. *Advances in Neural Information Processing Systems*, pages 1028–1036, 2016.

C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

T. P. Minka. Expectation propagation for approximate Bayesian inference. *Uncertainty in Artificial Intelligence*, 2001.

S. Barthelmé and N. Chopin. ABC-EP: Expectation propagation for likelihood-free Bayesian computation. *Internetional Conference on Machine Learning*, pages 289–296, 2011.

S. Tavaré, D. J. Balding, R. C. Griffiths, and P. Donnelly. Inferring coalescence times from DNA sequence data. *Genetics*, 145(2):505–518, 1997.

J. K. Pritchard, M. T. Seielstad, A. Perez-Lezaun, and M. W. Feldman. Population growth of human y chromosomes: a study of Y chromosome microsatellites. *Molecular Biology and Evolution*, 16(12):1791–1798, 1999.

M. N. Tran, D. J. Nott, and R. Kohn. Variational Bayes with intractable likelihood. *Journal of Computational and Graphical Statistics*, 26(4):873–882, 2017a.

Jiaxin S., Shengyang S., and Jun Z. Kernel implicit variational inference. *International Conference on Learning Representations*, 2018.

M. G. B. Blum and O. François. Non-linear regression models for approximate Bayesian computation. *Statistics and Computing*, 20(1):63–73, 2010.

T. A. Le, A. G. Baydin, R. Zinkov, and F. Wood. Using synthetic data to train neural networks is model-based reasoning. *International Joint Conference on Neural Networks*, 2017b.

M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman. Synthetic data and artificial neural networks for natural scene text recognition. *arXiv preprint arXiv:1406.2227*, 2014.

M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman. Reading text in the wild with convolutional neural networks. *International Journal of Computer Vision*, 116(1):1–20, 2016.

A. Gupta, A. Vedaldi, and A. Zisserman. Synthetic data for text localisation in natural images. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2315–2324, 2016.

F. Stark, C. Hazırbas, R. Triebel, and D. Cremers. Captcha recognition with active deep learning. *Workshop New Challenges in Neural Computation*, page 94, 2015.

Y. Güçlütürk, U. Güçlü, R. van Lier, and M. A. J. van Gerven. Convolutional sketch inversion. *European Conference on Computer Vision*, pages 810–824, 2016.

L. Ambrogioni, U. Güçlü, E. Maris, and M. van Gerven. Estimating nonlinear dynamics with the ConvNet smoother. *arXiv preprint arXiv:1702.05243*, 2017a.

L. Mescheder, S. Nowozin, and A. Geiger. Adversarial variational Bayes: Unifying variational autoencoders and generative adversarial networks. *arXiv preprint arXiv:1701.04722*, 2017.

D. Tran, R. Ranganath, and David M. Blei. Hierarchical implicit models and likelihood-free variational inference. *arXiv preprint arXiv:1702.08896*, 2017b.

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014.

S. Arora, A. Risteski, and Y. Zhang. Do GANs learn the distribution? Some theory and empirics. *International Conference on Learning Representations*, 2018.

S. Arora, R. Ge, Y. Liang, T. Ma, and Y. Zhang. Generalization and equilibrium in generative adversarial nets (GANs). *arXiv preprint arXiv:1703.00573*, 2017.

K. Csilléry, M. G. B. Blum, O. E. Gaggiotti, and O. François. Approximate Bayesian computation (ABC) in practice. *Trends in Ecology & Evolution*, 25(7):410–418, 2010.

J. M. Marin, P. Pudlo, C. P. Robert, and R. J. Ryder. Approximate Bayesian computational methods. *Statistics and Computing*, 22(6):1167–1180, 2012.

K. P. Murphy. *Machine Learning, A Probabilistic Perspective*. The MIT press, 2012.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

M. West. *Bayesian Forecasting*. Wiley Online Library, 1996.

E. N. Lorenz. Deterministic nonperiodic flow. *Journal of the Atmospheric Sciences*, 20(2):130–141, 1963.

F. Yu and V. Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015.

L. Ambrogioni, U. Güçlü, M. A. J. van Gerven, and E. Maris. The kernel mixture network: A nonparametric method for conditional density estimation of continuous random variables. *arXiv preprint arXiv:1705.07111*, 2017b.

G. Evensen. *Data Assimilation: The Ensemble Kalman Filter*. Springer, 2009.

T. M. Mitchell. *Machine Learning*. McGraw Hill, 1997.

S. Tokui, K. Oono, S. Hido, and J. Clayton. Chainer: A next-generation open source framework for deep learning. *Workshop on Machine Learning Systems (NIPS)*, 5, 2015.

David. Harrison and D. L. Rubinfeld. Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management*, 5(1): 81–102, 1978.

B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *The Annals of Statistics*, 32 (2):407–499, 2004.

W. N. Street, W. H. Wolberg, and O. L. Mangasarian. Nuclear feature extraction for breast tumor diagnosis. *Biomedical Image Processing and Biomedical Visualization*, 1905:861–871, 1993.

Y. Freund, R. Schapire, and N. Abe. A short introduction to boosting. *Journal of the Japanese Society For Artificial Intelligence*, 14(771-780):1612, 1999.

L. Breiman. Random forests. *Machine Learning*, 45 (1):5–32, 2001.

J. R. Quinlan. Induction of decision trees. *Machine Learning*, 1(1):81–106, 1986.

Thomas G Dietterich. Ensemble methods in machine learning. *International Workshop on Multiple Classifier Systems*, pages 1–15, 2000.