Submission and Formatting Instructions for International Conference on Machine Learning (ICML 2016)

Abstract

The purpose of this document is to provide both the basic paper template and submission guidelines. Abstracts should be a single paragraph, between 4–6 sentences long, ideally. Gross violations will trigger corrections at the camera-ready phase.

1. Introduction

We approach the problem of capturing the statistical dependence structure of the conditionally independent variational factors in a flexible, Bayesian nonparametric manner. We augment

From the perspective of the variational principle, we propose a drop-in for mean field 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 hierarchical variational distribution (CITE RANAGNATH 2015)... integrating over the auxiliary inference parameters recovers a marginal distribution as an infinite mixture of augmented distributions...

From the perspective of 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???)

The marginal variational distribution over the variables of interest, $q(Z, \eta)$, is itself intractable precluding the analytticity of the variational entropy... 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.

l;asdf. 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.
- Black-box approaches and SVI. Reparam trick for variance reduction, autoencoding and back-constraints for amortization.
- Augmented inference space, hierarchical variational distribution.

2.2. The Bayesian GP-LVM

3. GP-LVM-Conditional Mean Field Variational Inference

————-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.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

- 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

$$\begin{aligned}\log p_\theta(Y) &= \mathbb{E}_{q_\phi(Z)}[\log p_\theta(Y|Z) - \log q_\phi(Z)] \\ &\quad + \mathrm{KL}[q_\phi(Z)\|p_\theta(Z|Y)] \qquad (1)\\ &\geq \mathbb{E}_{q_\phi(Z)}[\log p_\theta(Y|Z) - \log q_\phi(Z)] =: \mathcal{L}_1 \qquad (2)\end{aligned}$$

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 2 with respect to the auxiliary variables, noting that the marginal variational distribution over Z can be expressed as $q(Z) = q(Z|F(), U, X)q(F(), U, X)/q(F(), U, X|Z)$, and introducing an auxiliary model $r_\psi(F(), U, X)$, we arrive at the following *auxiliary lower bound*:

$$\begin{aligned}\mathcal{L}_1 &\geq \mathbb{E}_{q(Z,F(),U,X)}\big[\log p(Y|Z)p(Z)r(F(), U, X|Z) \\ &\quad - \log q(Z|F(), X)q(F(), U, X)\big] \qquad (3)\\ &\quad + \mathbb{E}_{q(Z)}\big[\mathrm{KL}[q(F(), U, X|Z)\|r(F(), U, X|Z)]\big] =: \mathcal{L}_{aux}. \qquad (4)\end{aligned}$$

To explain this bound, we find it useful to refer to $q(F(), U, X)$ (resp. $q(F(), U, X|Z)$) 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

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:

$$\sum_{j=1}^{J}[\log p(Y|Z^{(j)})p(Z^{(j)})...] \qquad (5)$$

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 (5) will suffer from prohibitively high variance...

variance reduction ... reparam ... A but we also note that the additional lower bounding step (3) allows the resulting auxiliary lower bound (4) to be written as

We can rearrange (4) as follows:

$$\begin{aligned}\mathcal{L}_{aux} &= \mathbb{E}_{q(Z)}[\log p(Y|Z)] - \mathbb{E}_{q(F(),U,X)}[\mathrm{KL}[q(Z|F(), U, X)\|p(Z)]] \\ &\quad - \mathbb{E}_{q(Z)}[\mathrm{KL}[q(F(), U, X)\|r(F(), U, X|Z)]]. \qquad (6)\end{aligned}$$

As noted by Kingma and Welling, in the case of Gaussian $p(Z)$ and $q(Z|F(), U, X)$, 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 $r(F()|U, X, Z) := q(F()|U, X)$ which means that $r(F(), U, X|Z) = q(F()|U, X)r(U, X|Z)$. 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)...