On Sparse Variational Methods and the Kullback-Leibler Divergence between Stochastic Processes

Alexander G. de G. Matthews[1], **James Hensman**[2], **Richard E. Turner**[1], **Zoubin Ghahramani**[1]

[1]University of Cambridge, [2]Lancaster University

Abstract

The variational framework for learning inducing variables (Titsias, 2009a) has had a large impact on the Gaussian process literature. The framework may be interpreted as minimizing a rigorously defined Kullback-Leibler divergence between the approximating and posterior processes. To our knowledge this connection has thus far gone unremarked in the literature. In this paper we give a substantial generalization of the literature on this topic. We give a new proof of the result for infinite index sets which allows inducing points that are not data points and likelihoods that depend on all function values. We then discuss augmented index sets and show that, contrary to previous works, marginal consistency of augmentation is not enough to guarantee consistency of variational inference with the original model. We then characterize an extra condition where such a guarantee is obtainable. Finally we show how our framework sheds light on interdomain sparse approximations and sparse approximations for Cox processes.

1 Introduction

The variational approach to inducing point selection of Titsias (2009a) has been highly influential in the active research area of scalable Gaussian process approximations. The chief advantage of this particular framework is that the inducing points positions are variational parameters rather than model parameters and as such are protected from overfitting. In this paper we argue that whilst this is true, it may not be for exactly the reasons previously thought. The original framework is applied to conjugate likelihoods and has been extended to non-conjugate likelihoods (Chai, 2012; Hensman et al., 2015). An important advance in the use of variational methods was their combination with stochastic gradient descent (Hoffman et al., 2013) and the variational inducing point framework has been combined with such methods in the conjugate (Hensman et al., 2013) and non-conjugate cases (Hensman et al., 2015). The approach has also been successfully used to perform scalable inference in more complex models such as the Gaussian process latent variable model (Titsias and Lawrence, 2010; Damianou et al., 2015) and the related Deep Gaussian process (Damianou and Lawrence, 2013; Hensman and Lawrence, 2014).

To be more concrete let us set up some notation. Consider a function f mapping an index set X to the set of real numbers $f : X \mapsto \mathbb{R}$. Entirely equivalently we may write $f \in \mathbb{R}^X$ or use sequence notation $(f(x))_{x\in X}$. We also define set indexing of the function. If $S \subseteq X$ is some subset of the index set, then $f_S := (f(x))_{x\in S}$. We can put this notation to immediate use by defining a subset $D \subseteq X$ of the index set, of size N, that corresponds to those input points for which we have observed data. The corresponding function values will then be denoted f_D. For simplicity, we will initially assume that we have one, possibly noisy, possibly non-conjugate observation y per input data point which will together form a set Y.

Gaussian processes allow us to define a prior over functions f. After we observe the data we will have some posterior which we wish to approximate with a sparse distribution. At the heart of the variational induc-

Appearing in Proceedings of the 19^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2016, Cadiz, Spain. JMLR: W&CP volume 41. Copyright 2016 by the authors.

ing point approximation is the idea of 'augmentation' that appears in the original paper and many subsequent ones. We choose to monitor a set $Z \subseteq X$ of size M. These points may have some overlap with the input data points D but to give a computational speed up M will need to be less than the number of data points N. The Kullback-Leibler divergence given as an optimization criterion in Titsias' original paper is

$$\begin{aligned}&\mathcal{KL}[q(f_{D\setminus Z}, f_Z)||p(f_{D\setminus Z}, f_Z|Y)]\\ &= \int q(f_{D\setminus Z}, f_Z) \log\left\{\frac{q(f_{D\setminus Z}, f_Z)}{p(f_{D\setminus Z}, f_Z|Y)}\right\} df_{D\setminus Z} df_Z\,.\end{aligned} \tag{1}$$

The variational distribution at those data points which are not also inducing points is taken to have the form:

$$q(f_{D\setminus Z}, f_Z) := p(f_{D\setminus Z}|f_Z)q(f_Z) \tag{2}$$

where $p(f_{D\setminus Z}|f_Z)$ is the prior conditional and $q(f_Z)$ is a variational distribution on the inducing points only. Under this factorization, for a conjugate likelihood, the optimal $q(f_Z)$ has an analytic Gaussian solution (Titsias, 2009a). The non-conjugate case was then studied in subsequent work (Chai, 2012; Hensman et al., 2015). In both cases the sparse approximation requires only $\mathcal{O}(NM^2)$ rather than the $\mathcal{O}(N^3)$ required by exact methods in the conjugate case, or many commonly used non-conjugate approximations that don't assume sparsity.

The augmentation is justified by arguing that the model remains marginally the same when the inducing points are added. It is therefore suggested that variational inference in the augmented model, including for the parameters of said augmentation, is equivalent to variational inference in the original model, i.e that the inducing point positions can be considered to be variational parameters and are consequently protected from overfitting. For example see Titsias' original conference paper (Titsias, 2009a), section 3 or the longer technical report version (Titsias, 2009b), section 3.1. In the common case in the literature where the argument proceeds by applying Jensen's inequality to the marginal likelihood as, for example, in Hensman et al (2015) equations (6) and (17), the slack of the bound on the marginal likelihood is precisely the $\mathcal{KL}$-divergence (1). Therefore maximizing such a bound is exactly equivalent to minimizing this objective and the considerations that follow all apply.

In fact in this paper, whilst we applaud the excellent prior work, we will show that variational inference in an augmented model is not equivalent to variational inference in the original model. Without this justification, the $\mathcal{KL}$-divergence in equation (1) could seem to be a strange optimization target. The $\mathcal{KL}$-divergence has the inducing variables on both sides, so it might seem that in optimizing the inducing point positions we are trying to hit a 'moving target'. It is desirable to rigorously formulate a 'one sided' $\mathcal{KL}$-divergence that leads to Titsias' formulation. Such a derivation could be viewed as putting these elegant and popular methods on a firmer foundation, and is the topic of this article. As we shall show, this cements the framework for sparse interdomain inducing approximations and sparse variational inference in Cox processes. We wish to re-emphasize our respect for the previous work and for the avoidance of suspense we will find that much of the existing work carries over *mutatis mutandis*. Nevertheless we feel that most readers at the end of the paper will agree that a precise treatment of the topic should be of benefit going forward.

In terms of prior work for the theoretical aspect, the major other references are the early work of Seeger (2003a; 2003b). In particular Seeger identifies the $\mathcal{KL}$-divergence *between processes* (more commonly referred to as a relative entropy in those texts) as a measure of similarity and applies it to PAC-Bayes and to subset of data sparse methods. Crucially, Seeger outlines the rigorous formulation of such a $\mathcal{KL}$-divergence which is a large technical obstacle. Here we give a shorter, more general, and intuitive proof of the key theorem. We extend the stochastic process formulation to inducing points which are not necessarily selected from the data and show that this is equivalent to Titsias' formulation. In so far as we are aware this relationship has not previously been noted in the literature. The idea of using the $\mathcal{KL}$-divergence between processes is also mentioned in the early work of Csato and Opper (2002; 2002) but the transition from finite dimensional multivariate Gaussians to infinite dimensional Gaussian processes is not covered at the level of detail discussed here. An optimization target that in intent seems to be similar to a $\mathcal{KL}$-divergence between stochastic process is briefly mentioned in the work of Alvarez (2011). The notation used suggests that the integration is with respect to an 'infinite dimensional Lebesgue measure', which as we shall see is an argument that arrives at the right answer via a mathematically flawed route. Chai (2012) seems to have been at least partly aware of Seeger's $\mathcal{KL}$-divergence theorems (Seeger, 2003b) but instead uses them to bound the finite joint predictive probability of a non sparse process.

This article proceeds by first discussing the finite dimensional version of the full argument. This requires considerably less mathematical machinery and much of the intuition can be gained from this case. We then proceed to give the full measure theoretic formulation, giving a new proof that allows inducing points that

are not data points and for the likelihood to depend on infinitely many function values. Next we discuss augmentation of the original index set, using the crucial chain rule for $\mathcal{KL}$-divergences. This gives us a framework to discuss marginal consistency and how variational inference in augmented models is not necessarily equivalent to variational inference in the original model. We then show that under very general conditions augmentation which is deterministic conditioned on the whole latent function does have the desired property. We apply our results to sparse variational interdomain approximations and to posterior inference in Cox processes. Finally we conclude and highlight avenues for further research.

2 Finite index set case

This section is in fact a less general case of what follows. It is included for the benefit of those familiar with the previous work on variational sparse approximations and as an important special case. Consider the case where X is finite. We introduce a new set $* := X \backslash (D \cup Z)$, in words: all points that are in the index set that aren't inducing points or data points. These points might be of practical interest for instance when making predictions on hold out data.

We extend the variational distribution to include these points:

$$q(f_*, f_{D\backslash Z}, f_Z) := p(f_*, f_{D\backslash Z}|f_Z)q(f_Z). \tag{3}$$

We then consider the $\mathcal{KL}$-divergence between this extended variational distribution and the full posterior distribution $p(f|Y)$

$$\begin{aligned}
&\mathcal{KL}[q(f_*, f_{D\backslash Z}, f_Z)||p(f|Y)] \\
=&\mathcal{KL}[q(f_*, f_{D\backslash Z}, f_Z)||p(f_*, f_{D\backslash Z}, f_Z|Y)] \\
=&\int q(f_*, f_{D\backslash Z}, f_Z) \log \frac{q(f_*, f_{D\backslash Z}, f_Z)}{p(f_*, f_{D\backslash Z}, f_Z|Y)} df_* df_{D\backslash Z} df_Z
\end{aligned} \tag{4}$$

Next we expand the term inside the logarithm and cancel one of the terms that appears in both the numerator and the denominator:

$$\begin{aligned}
&\frac{q(f_*, f_{D\backslash Z}, f_Z)}{p(f_*, f_{D\backslash Z}, f_Z|Y)} \\
=&\frac{p(f_*|f_{D\backslash Z}, f_Z)p(f_{D\backslash Z}|f_Z)q(f_Z)p(Y)}{p(f_*|f_{D\backslash Z}, f_Z)p(f_{D\backslash Z}|f_Z)p(f_Z)p(Y|f_D)} \\
=&\frac{p(f_{D\backslash Z}|f_Z)q(f_Z)p(Y)}{p(f_{D\backslash Z}|f_Z)p(f_Z)p(Y|f_D)} \\
=&\frac{q(f_{D\backslash Z}, f_Z)}{p(f_{D\backslash Z}, f_Z|Y)}
\end{aligned} \tag{5}$$

Substituting back into the full integral and exploiting the marginalization property of the conditional density we obtain:

$$\begin{aligned}
&\int p(f_*, f_{D\backslash Z}|f_Z)q(f_Z) \log \frac{q(f_{D\backslash Z}, f_Z)}{p(f_{D\backslash Z}, f_Z|Y)} df_* df_{D\backslash Z} df_Z \\
=&\int p(f_{D\backslash Z}|f_Z)q(f_Z) \log \frac{q(f_{D\backslash Z}, f_Z)}{p(f_{D\backslash Z}, f_Z|Y)} df_{D\backslash Z} df_Z
\end{aligned} \tag{6}$$

The last line is exactly the $\mathcal{KL}$-divergence used by Titsias (2009a) that we already described in equation (1). We thus see that for finite index sets considering the $\mathcal{KL}$-divergence between the two distributions is equivalent to Titsias' $\mathcal{KL}$-divergence. We might choose to optimize our choice of the M by selecting them from the $|X|$ possible values in the index set and comparing the $\mathcal{KL}$-divergence between distributions given in equation (4). The equivalence with equation (1) that we have just derived shows us that in this case the appearance of the inducing values on both sides of the equation is just a question of 'accounting'. That is to say, whilst we are in fact optimizing the $\mathcal{KL}$-divergence between the full distributions, we only need to keep track of the distribution over function values f_Z and $f_{D\backslash Z}$. All the other function values f_* marginalize. For different choices of inducing points we will need to keep track of different function values and be able to safely ignore different values f_*.

3 Infinite index set case

3.1 There is no useful infinite dimensional Lebesgue measure

One might hope to cope with not only finite index sets but also infinite index sets in the way discussed in section 2. Unfortunately when X and hence f_* are infinite sets we cannot integrate with respect to a 'infinite dimensional vector'. That is to say the notation $\int (\cdot) df_*$ can no longer be correctly used.

For a discussion of this see, for example, Hunt et al (1992). The crux of the issue is that to give sensible answers such a measure would need to be translation invariant and locally finite. Unfortunately the only measure that obeys these two properties is the zero measure which assigns zero to every input set. Thus we see that it will be necessary to rethink our approach to a $\mathcal{KL}$-divergence between stochastic processes. It will turn out that a reasonable definition will require the full apparatus of measure theory. Readers looking for some background on these issues may wish to consult a larger text (Billingsley, 1995; Capinski and Kopp, 2004).

3.2 The $\mathcal{KL}$-divergence between processes

In this section we review the rigorous definition of the $\mathcal{KL}$-divergence between stochastic processes (Gray, 2011).

Suppose we have two measures μ and η for (Ω, Σ) and that μ is absolutely continuous with respect to η. Then there exists a Radon-Nikodym derivative $\frac{d\mu}{d\eta}$ and the correct definition for $\mathcal{KL}$-divergence between these measures is:

$$\mathcal{KL}[\mu||\eta] = \int_\Omega \log\left\{\frac{d\mu}{d\eta}\right\} d\mu\,. \tag{7}$$

In the case where μ is not absolutely continuous with respect to η we let $\mathcal{KL}[\mu||\eta] = \infty$. In the case where the sample space is $\mathbb{R}^K$ for some finite K and both measures are dominated by Lebesgue measure m this reduces to the more familiar definition:

$$\mathcal{KL}[\mu||\eta] = \int_\Omega u \log\left\{\frac{u}{v}\right\} dm \tag{8}$$

where u and v are the respective densities with respect to Lebesgue measure. The first definition is more general and allows us to deal with the problem of there being no sensible infinite dimensional Lebesgue measure by instead integrating with respect to the measure μ.

3.3 A general derivation of the sparse inducing point framework

In this section we give a general derivation of the sparse inducing point framework. The derivation is more general than that of Seeger (2003a; 2003b) since it does not require that the inducing points are selected from the data points. Nor does it assume that the relevant finite dimensional marginal distributions have density with respect to Lebesgue measure. Finally since the dependence on the elegant properties of Radon-Nikodym derivatives has been made more explicit we believe it is clearer *why* the derivation works and how one would generalize it.

We are now interested in three types of probability measure on sets of functions $f : X \mapsto \mathbb{R}$. The first is the prior measure P which will be assumed to be a Gaussian process. The second is the approximating measure Q which will be assumed to be a sparse Gaussian process and the third is the posterior process $\hat{P}$ which may be Gaussian or non-Gaussian depending on whether we have a conjugate likelihood. We start with a measure theoretic definition of Bayes' theorem for a dominated model (Schervish, 1995). It specifies the Radon-Nikodym derivative of the posterior with respect to the prior.

$$\frac{d\hat{P}}{dP}(f) = \frac{L(Y|f)}{L(Y)} \tag{9}$$

with $L(Y|f)$ being the likelihood and $L(Y) = \int_{\mathbb{R}^X} L(Y|f)dP(f)$ the marginal likelihood. As we have assumed in previous sections we will initially restrict the likelihood to only depend on the finite data subset of the index set. We denote by $\pi_C : \mathbb{R}^X \mapsto \mathbb{R}^C$ a projection function, which takes the whole function as an argument and returns the function at some set of points C. In this case we have:

$$\frac{d\hat{P}}{dP}(f) = \frac{d\hat{P}_D}{dP_D}(\pi_D(f)) = \frac{L(Y|\pi_D(f))}{L(Y)} \tag{10}$$

and similarly the marginal likelihood only depends on the function values on the data set $L(Y) = \int_{\mathbb{R}^D} L(Y|f_D)dP_D(f_D)$. In fact, we will relax the assumption that the data set is finite in section 5.2 and the ability to do so is one of the benefits of this framework. Next we specify Q by assuming it has density with respect to the posterior and thus the prior and that the density with respect to the prior depends on some set of points Z:

$$\frac{dQ}{dP}(f) = \frac{dQ_Z}{dP_Z}(\pi_Z(f))\,. \tag{11}$$

Under this assumption Q is fully specified if we know P and $\frac{dQ_Z}{dP_Z}$. To gain some intuition for this assumption we can compare equations (11) and (10). We see that in the approximating distribution the set Z is playing a similar one to that played for D in the true posterior distribution. We now bring these assumptions together. Let us apply the chain rule for Radon-Nikodym derivatives and a standard property of logarithms:

$$\begin{aligned}&\mathcal{KL}[Q||\hat{P}]\\ &= \int_{\mathbb{R}^X} \log\left\{\frac{dQ}{dP}(f)\right\} dQ(f) - \int_{\mathbb{R}^X} \log\left\{\frac{d\hat{P}}{dP}(f)\right\} dQ(f)\,.\end{aligned} \tag{12}$$

Taking the first term alone we exploit the sparsity assumption for the approximating distribution:

$$\begin{aligned}&\int_{\mathbb{R}^X} \log\left\{\frac{dQ}{dP}(f)\right\} dQ(f)\\ &= \int_{\mathbb{R}^Z} \log\left\{\frac{dQ_Z}{dP_Z}(f_Z)\right\} dQ_Z(f_Z)\,.\end{aligned} \tag{13}$$

Taking the second term in the last line of equation (12) and exploiting the measure theoretic Bayes' theorem

we obtain:

$$\begin{aligned}&\int_{\mathbb{R}^X} \log\left\{\frac{d\hat{P}}{dP}(f)\right\} dQ(f)\\ =&\int_{\mathbb{R}^D} \log\left\{\frac{d\hat{P}_D}{dP_D}(f_D)\right\} dQ_D(f_D)\\ =&\mathbb{E}_{Q_D}\left[\log L(Y|f_D)\right] - \log L(Y)\,. \end{aligned} \tag{14}$$

Finally noting the appearance of a marginal $\mathcal{KL}$-divergence we obtain our result:

$$\begin{aligned}\mathcal{KL}[Q||\hat{P}] = \mathcal{KL}[Q_Z||P_Z] - \mathbb{E}_{Q_D}[\log L(Y|f_D)]\\ + \log L(Y)\,. \end{aligned} \tag{15}$$

As is common with variational approximations, in most cases of interest the marginal likelihood will be intractable. However since it is an additive constant, independent of Q, it can be safely ignored. The final equation shows that we need to be able to compute the $\mathcal{KL}$-divergence between the inducing point marginals of the approximating distribution and the prior for all $Z \subset X$ and the expectation under the data marginal distribution of Q of the log likelihood. In the case where the likelihood factorizes across data terms this will give a sum of one dimensional expectations. Note the similarity of equation (15) with Hensman et al. (2015) equation (17) where a less general expression is motivated from a 'model augmentation' view. Notice that at no point in our derivation did we try to invoke the pathological 'infinite dimensional Lebesgue measure' which is important for the reasons discussed in section 3.1. The ease of derivation suggests that Radon-Nikodym derivatives and measure theory provide the most natural and general way to think about such approximations.

4 Augmented index sets

We now consider the case where we supplement the original (finite or infinite) index set X with a finite set of elements I, intending to use them as inducing points. The precise nature of the augmented prior model will be parameterized by some parameters θ which we will hope to tune to give a good approximation. It will be seen that the this is very much in the spirit of the original augmentation argument given by Titsias (2009a) and the 'variational compression' framework of Hensman and Lawrence (2014). This setup also covers the case of variational 'interdomain' Gaussian processes which were mooted but not implemented in Figueiras-Vidal and Lazaro-Gredilla (2009) and implemented under the basis of the marginal consistency argument in Alvarez et al (2011). We intend to discuss the marginal consistency argument in some detail and we shall deal with the thorny issues surrounding the rigorous treatment of the various infinities involved.

Marginal consistency is easily ensured by specifying the distribution of the augmented function value points f_I conditioned on the values of the function on the original set f_X. We denote the corresponding measure as $P_{I|X}(\bullet\,;\theta)$[1]. Let $\Omega_X = \mathbb{R}^X$ and $\Omega_I = \mathbb{R}^I$ be the sample spaces associated with the original index set and the augmenting variables respectively. Let $\mathcal{F}_X$ and $\mathcal{F}_I$ be their σ-algebras. Marginal consistency states that we will be interested in probability measures that have the following behaviour on the measurable rectangles $A_X \times A_I \in \mathcal{F}_X \times \mathcal{F}_I$:

$$P_{X\cup I}(A_X \times A_I;\theta) = \int_{A_X} P_{I|X}(A_I;\theta) dP_X(f_X). \tag{16}$$

We have included the augmentation parameters θ explicitly up until now, but for brevity we will omit them in what follows. We will make this marginal consistency assumption in all that follows. Let us call the overall set $X \cup I$ the 'union set'. In a similar vein to the previous section we assume that the approximating measure $Q_{X\cup I}$ has density with respect to the augmented prior model $P_{X\cup I}$ and that the Radon-Nikodym derivative is only a function of the augmented function points:

$$\frac{dQ_{X\cup I}}{dP_{X\cup I}}(f_{X\cup I}) = \frac{dQ_I}{dP_I}(\pi_I(f_{X\cup I}))\,. \tag{17}$$

Acting as if the augmented set were the original index set we would obtain by a similar argument:

$$\begin{aligned}\mathcal{KL}[Q_{X\cup I}||\hat{P}_{X\cup I}] = \mathcal{KL}[Q_I||P_I] - \mathbb{E}_{Q_D}[\log L(Y|f_D)]\\ + \log L(Y)\,. \end{aligned} \tag{18}$$

Sharp eyed readers, however, will have noted that since $\hat{P}_{X\cup I}$ depends on the augmentation parameters θ we are back in a situation where we can tune the approximation on the left hand side and the optimization target on the right. As we will see in the next section we are not necessarily rescued by the marginal consistency argument. It is not the case in general that $\mathcal{KL}[Q_X||\hat{P}_X]$ equals $\mathcal{KL}[Q_{X\cup I}||\hat{P}_{X\cup I}]$. In fact the relationship is governed by the chain rule for $\mathcal{KL}$-divergences as we shall now see.

4.1 The chain rule for $\mathcal{KL}$-divergences

For what follows we will require the chain rule for $\mathcal{KL}$-divergences (Gray, 2011). Let U and V be two Polish

[1] Note that for brevity our notation for conditional measures won't include the explicit function dependence. For example, in this case we omit the explicit dependence on f_X.

spaces endowed with their standard Borel σ-algebras and let $U \times V$ be the Cartesian product of these spaces endowed with the corresponding product σ-algebra. Consider two probability measures $\mu_{U\times V}, \eta_{U\times V}$ on this product space and let $\mu_{U|V}, \eta_{U|V}$ be the corresponding regular conditional measures. Assume that $\mu_{U\times V}$ is dominated by $\eta_{U\times V}$. The chain rule for $\mathcal{KL}$-divergences says that:

$$\mathcal{KL}[\mu_{U\times V}||\eta_{U\times V}] = \mathbb{E}_{\mu_V}\{\mathcal{KL}[\mu_{U|V}||\eta_{U|V}]\} + \mathcal{KL}[\mu_V||\eta_V]\,. \tag{19}$$

The first term on the right hand side is referred to as the 'conditional $\mathcal{KL}$-divergence' or 'conditional relative entropy'.

4.2 The marginally consistent augmentation argument is not correct in general.

Applying the chain rule for $\mathcal{KL}$-divergences to the divergence on the union set we obtain:

$$\begin{aligned}&\mathcal{KL}[Q_{X\cup I}||\hat{P}_{X\cup I}]\\ &= \mathbb{E}_{Q_X}\left\{\mathcal{KL}[\,Q_{I|X}||\hat{P}_{I|X}]\,\right\} + \mathcal{KL}[Q_X||\hat{P}_X]\\ &= \mathbb{E}_{Q_X}\left\{\mathcal{KL}[\,Q_{I|X}||P_{I|X}]\,\right\} + \mathcal{KL}[Q_X||\hat{P}_X]\,.\end{aligned} \tag{20}$$

The final line follows from the fact that in the assumed model augmentation scheme the additional variables f_I are conditionally independent of the data given f_X. This relation makes precise our claim that marginal consistency is not enough to guarantee that $\mathcal{KL}[Q_X||\hat{P}_X]$ equals $\mathcal{KL}[Q_{X\cup I}||\hat{P}_{X\cup I}]$. In fact this will only be true if $Q_{I|X} = P_{I|X}$, Q_X-almost surely. In the case where this is not true variational inference in the family of augmented models is not equivalent to variational inference in the original model and we will be optimizing a 'two-sided' objective function. We will consider an important condition which ensures the desired equality does hold in the next section.

Before we move on, however, it is also instructive to consider a transformation of the original unaugmented problem into the augmented problem. Take the transformed augmentation set and index set $(\tilde{I}, \tilde{X})$ to be defined in terms of the old sets as $(X\backslash D, D)$. The chain rule then tells us that the $\mathcal{KL}$-divergence on the data set is not in general equal to the $\mathcal{KL}$-divergence on the index set although this is true if $Z \subset D$.

4.3 Deterministic augmentation

Here we discuss an important case where the augmented $\mathcal{KL}$-divergence and the unaugmented $\mathcal{KL}$-divergence are indeed equal, namely where the additional variables f_I are a deterministic function h of the function values on the original index set f_X. A few conceptual points may be useful before we go into the detail. First the constraint only says that the values are deterministic conditioned on the function over the whole index set and the index set itself may be infinite. Usually in practice either through noise, finite observations or both, we can't know the latent function exactly and hence in our model we won't know the inducing variables exactly. Second, whilst this assumption may initially seem contrived, in fact it covers two very important cases: the original framework where some inducing points are selected from the index set X then 'copied' over to I and as we shall see later the interdomain inducing point framework. Having a deterministic function mapping is equivalent to having a delta function conditional distribution centred on the function value. Thus the conditional $\mathcal{KL}$-divergence term in equation (20) i.e the expectation of the conditional on the right hand side, will be zero if the approximating measure $Q_{X\cup I}$ has the same delta function conditional. The next theorem shows that if we follow the usual prescription for defining $Q_{X\cup I}$ this will indeed be the case.

4.3.1 The governing theorem on deterministic augmentation

Let $(\Omega_X, \mathcal{F}_X)$ and $(\Omega_I, \mathcal{F}_I)$ be two Polish spaces and let $(\Omega_X \times \Omega_I, \mathcal{F}_X \times \mathcal{F}_I)$ be their product space endowed with product σ-algebra. Let $h : \Omega_X \mapsto \Omega_I$ be a $\mathcal{F}_X/\mathcal{F}_I$ measurable function. We are interested in a measure $P : \mathcal{F}_X \times \mathcal{F}_I \mapsto \mathbb{R}$ which has the following property on the measurable rectangles $A_X \times A_I$

$$P(A_X \times A_I) = P_X(A_X \cap h^{-1}(A_I)) \tag{21}$$

where $P_X := P(A_X \times \Omega_I)$ is the marginal distribution for X. This assumption in turn implies that the marginal distribution for I has the form

$$P_I(A_I) = P_X(h^{-1}(A_I)) \tag{22}$$

which is the *push forward measure* of P_X under the function h. It is clear that the regular conditional distribution $P_{I|X}(\bullet)$ has a point measure property:

$$P_{I|X}(A_I) = \delta_{h(f_X)}(A_I)\,. \tag{23}$$

Let $P_{X|I}(\bullet)$ be the regular conditional distribution of f_X conditioned on f_I. Next we define a second measure $Q : \mathcal{F}_X \times \mathcal{F}_I \mapsto \mathbb{R}$ which has the following property on measurable rectangles

$$Q(A_X \times A_I) = \int_{A_I} P_{X|I}(A_X) dQ_I(f_I)\,. \tag{24}$$

Finally we assume that $Q_I << P_I$. The theorem states that under the assumptions of the previous section the

marginal distributions of Q have the following property:

$$Q_I(A_I) = Q_X(h^{-1}(A_I)) \,. \tag{25}$$

That is to say the marginal distribution of Q for Z is the push forward measure of Q_X under the function h. Consequently the approximating distribution for f_I conditioned on f_X also has the point measure property

$$Q_{I|X}(A_I) = \delta_{h(f_X)}(A_I) \,. \tag{26}$$

We now give a proof. Starting from the right hand side of equation (25)

$$\begin{aligned} Q_X(h^{-1}(A_I)) &= Q(h^{-1}(A_I) \times \Omega_I) \\ &= \int_{\Omega_I} P_{X|I}(h^{-1}(A_I)) dQ_I(f_I) \,. \end{aligned} \tag{27}$$

Next since $Q_I << P_I$ we apply the Radon-Nikodym theorem:

$$\begin{aligned} &\int_{\Omega_I} P_{X|I}(h^{-1}(A_I)) dQ_I(f_I) \\ =&\int_{\Omega_I} P_{X|I}(A_X) \frac{dQ_I}{dP_I} dP_I(f_I) \,. \end{aligned} \tag{28}$$

The existence of conditional distributions is also guaranteed by the Radon-Nikodym theorem. Explicitly we have

$$P_{X|I}(A_X) = \frac{dP(A_X \times \bullet)}{dP_I(\bullet)} \,. \tag{29}$$

Continuing on from equation (28) and applying an elementary theorem of Radon-Nikodym derivatives we have:

$$\begin{aligned} &\int_{\Omega_I} P_{X|I}(h^{-1}(A_I)) \frac{dQ_I}{dP_I} dP_I(f_I) \\ =&\int_{\Omega_I} \frac{dQ_I}{dP_I} dP(h^{-1}(A_I) \times f_I) \,. \end{aligned} \tag{30}$$

Now we apply the property given by equation (21)

$$\begin{aligned} &\int_{\Omega_I} \frac{dQ_I}{dP_I} dP(h^{-1}(A_I) \times f_I) \\ =&\int_{\Omega_I} \frac{dQ_I}{dP_I} dP_X(h^{-1}(A_I) \cap h^{-1}(f_I)) \,. \end{aligned} \tag{31}$$

Now we apply some algebraic manipulations of the integral:

$$\begin{aligned} &\int_{\Omega_I} \frac{dQ_I}{dP_I} dP_X(h^{-1}(A_I) \cap h^{-1}(f_I)) \\ =&\int_{\Omega_I} \frac{dQ_I}{dP_I} dP_X(h^{-1}(A_I \cap f_I)) \\ =&\int_{\Omega_I} \frac{dQ_I}{dP_I} dP_I(A_I \cap f_I) \\ =&\int_{A_I} \frac{dQ_I}{dP_I} dP_I(f_I) = Q_I(A_I) \end{aligned} \tag{32}$$

as was claimed.

5 Examples

5.1 Variational interdomain approximations

Here we consider the sparse variational interdomain approximation which was suggested but not realized in Figueiras-Vidal and Lazaro-Gredilla (2009) and appeared under the basis of the marginal consistency argument in Alvarez et al (2011). An interdomain variable is a random variable, indexed by $i \in I$ defined in the following way:

$$f_i(\theta) = \int_X g_i(x, \theta) f_x \, d\lambda(x) \tag{33}$$

Here λ is a measure on X with some appropriate σ-algebra, $\{g_i : i \in I\}$ is a set of λ-integrable functions from X to $\mathbb{R}$. The interdomain variables may be viewed as deterministic conditional on the whole function f_X so the theorems of section 4.3 come into play. Since the intention here is to put this framework on a firm logical footing, we should also consider the thorny issue of the measurability of this transformation and the associated random variable. The existence of separable, measurable, versions of stochastic processes, including most commonly used Gaussian processes, was settled in the work of Doob (1953). It also discusses the conditions necessary to apply Fubini's theorem to expectations of the random variable defined by equation (33). The application of Fubini's theorem is essential to the utility of such methods in practice (Figueiras-Vidal and Lázaro-Gredilla, 2009).

Thus we may correctly optimize the parameters θ of interdomain inducing points, safe in the knowledge that this decision is variationally protected from overfitting and optimizes a well defined $\mathcal{KL}$-divergence objective. The potential for a wide variety of improved sparse approximations in this direction is thus, in our opinion, significant.

5.2 Approximations to Cox process posteriors

In this section we relax the assumption that the data set D is finite, which is necessary to consider Gaussian process based Cox processes. One specific case of this model is considered by Lloyd et al (2015) under the marginal consistency motivation. A Gaussian process based Cox process has the following generative scheme:

$$\begin{aligned} f &\sim \mathcal{GP}(m, K) \\ h &= \rho(f) \\ Y|h &\sim \mathcal{PP}(h) \,. \end{aligned} \tag{34}$$

Here $\mathcal{GP}(m, K)$ denotes a Gaussian process with mean m and kernel K, $\rho : \mathbb{R} \mapsto (0, \infty)$ is an inverse link

function, $\mathcal{PP}(h)$ is a Poisson process with intensity h and D is a set of points in the original index set X. For example in a geographical spatial statistics application we might take X to be some bounded subset of $\mathbb{R}^2$. The key issue with the Poisson process likelihood is that it depends not just on those members of X where points were observed but in fact on all points in X. Intuitively the absence of points in an area suggests that the intensity is lower there. Thus $D = X$. The likelihood in question is:

$$L(Y|f_D) = \left(\prod_{y\in Y} \rho(y)\right) \exp\left\{-\int_X \rho(x)dm(x)\right\}. \tag{35}$$

where m denotes for instance Lebesgue measure on X. The full X dependence manifests itself through the integral on the right hand side. We will require that the integral exists almost surely. In Lloyd et al (2015) equation (3), the application of Bayes' theorem appears to require a density with respect to infinite dimensional Lebesgue measure. As pointed out in 3.1 such a notion is pathological. This however can be fixed because the more general form of Bayes' theorem in equation (9) of this paper still applies. Thus we can apply the results of section 3.3 to obtain:

$$\begin{aligned}\mathcal{KL}[Q||\hat{P}] =& \mathcal{KL}[Q_Z||P_Z] - \sum_{y\in Y} \mathbb{E}_{Q_y}\left[\log \rho(y)\right] \\ &+ \mathbb{E}_{Q_X}\left[\int_X \rho(x)dm(x)\right] + \log L(Y)\,.\end{aligned} \tag{36}$$

As in section 5.1 we will need to check that the conditions for Fubini's theorem apply (Doob, 1953) which gives:

$$\begin{aligned}\mathcal{KL}[Q||\hat{P}] =& \mathcal{KL}[Q_Z||P_Z] - \sum_{y\in Y} \mathbb{E}_{Q_y}\left[\log \rho(y)\right] \\ &+ \int_X \mathbb{E}_{Q_x}\left[\rho(x)\right] dm(x) + \log L(Y)\,.\end{aligned} \tag{37}$$

For the specific case of ρ used in Lloyd et al (2015) the working then continues as in that paper and the elegant results that follow all still apply. Note that one could combine these Cox process approximations with the interdomain framework and this could be a fruitful direction for further work.

6 Conclusion and acknowledgements

In this work we have elucidated the connection between the variational inducing point framework (Titsias, 2009a) and a rigorously defined $\mathcal{KL}$-divergence between stochastic processes. Early use of the rigorous formulation of $\mathcal{KL}$-divergence in the Gaussian processes for machine learning literature was made by Seeger (2003a; 2003b). Here we have increased the domain of applicability of those proofs by allowing for inducing points that are not data points, and removing unnecessary dependence on Lebesgue measure. We would argue that our proof clarifies the central and elegant role played by Radon-Nikodym derivatives. We then consider for the first time in this framework the case where additional variables are added solely for the purpose of variational inference. We show that marginal consistency is not enough to guarantee a principled optimization objective but that if we make the inducing points deterministic conditional on the whole function then a principled optimization objective is guaranteed and the parameters of the augmentation are variationally protected. We then show how the extended theory allows us to correctly handle principled interdomain sparse approximations and that we can cope correctly with the importance case of Cox processes where the likelihood depends on an infinite set of function points.

It seems reasonable to hope that elucidating the measure theoretic roots of the formulation will help the community to generalise the framework and lead to even better practical results. In particular it seems that since interdomain inducing points are linear functionals, the theory of Hilbert spaces might profitably be applied here. It also seems reasonable to think given the generality of section 3.3 that other Bayesian and Bayesian nonparametric models might be amenable to such a treatment.

The authors wish to thank Matthias Seeger, Michalis Titsias, Giles Shaw and the anonymous reviewer of a previous paper. AM and ZG would like to acknowledge EPSRC grant EP/I036575/1, and a Google Focused Research award. JH was supported by a MRC fellowship. RET thanks the EPSRC for funding (grant numbers EP/G050821/1 and EP/L000776/1).

References

Álvarez, M. A. (2011). *Convolved Gaussian process priors for multivariate regression with applications to dynamical systems*. PhD thesis, University of Manchester.

Álvarez, M. A. and Lawrence, N. D. (2011). Computationally Efficient Convolved Multiple Output Gaussian Processes. *J. Mach. Learn. Res.*, 12:1459–1500.

Billingsley, P. (1995). *Probability and Measure*. Wiley-Interscience, 3 edition.

Capinski, M. and Kopp, P. (2004). *Measure, Integral and Probability*. Springer Undergraduate Mathematics Series. Springer London.

Chai, K. M. A. (2012). Variational Multinomial Logit Gaussian Process. *J. Mach. Learn. Res.*, 13(1):1745–1808.

Csató, L. (2002). *Gaussian processes: iterative sparse approximations*. PhD thesis, Aston University.

Csató, L. and Opper, M. (2002). Sparse on-line Gaussian processes. *Neural computation*, 14(3):641–668.

Damianou, A. and Lawrence, N. (2013). Deep Gaussian processes. In Carvalho, C. and Ravikumar, P., editors, *Proceedings of the Sixteenth International Workshop on Artificial Intelligence and Statistics (AISTATS)*, AISTATS '13, pages 207–215. JMLR W&CP 31.

Damianou, A. C., Titsias, M. K., and Lawrence, N. D. (2015). Variational inference for latent variables and uncertain inputs in Gaussian processes. *Journal of Machine Learning Research (JMLR)*, 2.

Doob, J. (1953). *Stochastic Processes*. Wiley Publications in Statistics. John Wiley & Sons.

Figueiras-Vidal, A. and Lázaro-Gredilla, M. (2009). Inter-domain Gaussian processes for sparse inference using inducing features. In Bengio, Y., Schuurmans, D., Lafferty, J., Williams, C., and Culotta, A., editors, *Advances in Neural Information Processing Systems 22*, pages 1087–1095. Curran Associates, Inc.

Gray, R. M. (2011). *Entropy and Information Theory*. Springer-Verlag New York, Inc., New York, NY, USA, 2 edition.

Hensman, J., Fusi, N., and Lawrence, N. D. (2013). Gaussian processes for Big Data. In *Conference on Uncertainty in Artificial Intellegence*, pages 282–290. auai.org.

Hensman, J. and Lawrence, N. D. (2014). Nested Variational Compression in Deep Gaussian Processes. *ArXiv e-prints*.

Hensman, J., Matthews, A. G. d. G., and Ghahramani, Z. (2015). Scalable variational Gaussian process classification. In *18th International Conference on Artificial Intelligence and Statistics*, pages 351–360, San Diego, California, USA.

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

Hunt, B. R., Sauer, T., James, and Yorke, A. (1992). Prevalence: A translation-invariant almost every on infinite-dimensional spaces. *Bulletin of the Amer. Math. Soc*, pages 217–238.

Lloyd, C., Gunter, T., Osborne, M., and Roberts, S. (2015). Variational inference for Gaussian process modulated Poisson processes. In *Proceedings of The 32nd International Conference on Machine Learning*, pages 1814–1822.

Schervish, M. (1995). *Theory of Statistics*. Springer Series in Statistics. Springer.

Seeger, M. (2003a). *Bayesian Gaussian process models: PAC-Bayesian generalisation error bounds and sparse approximations*. PhD thesis, University of Edinburgh.

Seeger, M. (2003b). PAC-Bayesian generalisation error bounds for Gaussian process classification. *J. Mach. Learn. Res.*, 3:233–269.

Titsias, M. K. (2009a). Variational learning of inducing variables in sparse Gaussian processes. In *International Conference on Artificial Intelligence and Statistics 12*, pages 567–574.

Titsias, M. K. (2009b). Variational model selection for sparse Gaussian process regression. Technical report.

Titsias, M. K. and Lawrence, N. D. (2010). Bayesian Gaussian process latent variable model. In *Thirteenth International Conference on Artificial Intelligence and Statistics*.