Article:

Ward, W.O.C. orcid.org/0000-0002-4904-7294 and Álvarez, M.A. (Submitted: 2019) Variational bridge constructs for approximate Gaussian process regression. arXiv. (Submitted)

Variational Bridge Constructs for Approximate Gaussian Process Regression

Wil O C Ward
Department of Computer Science
Univesity of Sheffield
Sheffield, UK
w.ward@sheffield.ac.uk

Mauricio A Álvarez
Department of Computer Science
University of Sheffield
Sheffield, UK
mauricio.alvarez@sheffield.ac.uk

Abstract

This paper introduces a method to approximate Gaussian process regression by representing the problem as a stochastic differential equation and using variational inference to approximate solutions. The approximations are compared with full GP regression and generated paths are demonstrated to be indistinguishable from GP samples. We show that the approach extends easily to non-linear dynamics and discuss extensions to which the approach can be easily applied.

1 Introduction

Gaussian process (GP) regression is an effective tool for inferring the nature of a function given a small number of possibly noisy observations. However, in the case where these signals interact non-linearly, for example in non-linear multi-output GPs and latent force models (LFMs) (Hartikainen et al., 2012); or where the likelihoods are non-Gaussian (Hensman et al., 2015), approximation methods are necessary. Challenges lie in the effective estimation of the posterior in the case where it is intractible, as well as scalability for large datasets.

In the case of the former, there are a number of approaches including variational inference (VI). VI can also be utilised with sparse representations of the Gaussian approximation to deal with large datasets (Hensman et al., 2013). Alternatively, the GP can be represented as a (Markovian) stochastic differential equation (SDE) and inferred sequentially in linear time (Hartikainen and Särkkä, 2010). In more recent work, this approach has been adapted for non-Gaussian likelihoods (Nickisch et al., 2018).

The representation of GPs as SDEs has been used with covariance assumptions on, for example, smooth and (quasi-)periodic systems (Hartikainen and Särkkä, 2010; Solin and Särkkä, 2013). In these instances, the posterior can be inferred using Kalman filtering and smoothing (Särkkä, 2013); non-linear approximations thereof have been used for models with non-linear interaction, such as in the case of non-linear LFMs represented as SDEs (Hartikainen and Särkkä, 2011). An alternative approach to approximating non-linear SDEs was proposed by Ryder et al. (2018), constructing Brownian bridges conditioned on observations and using a variational distribution to represent the posterior solution to the SDE.

In this paper, we utilise these *variational bridges* to approximate Gaussian process regression in its SDE form, comparing the approximation against full GP regression. We demonstrate that the paths generated are effective approximations of a true GP, using maximum mean discrepancy metrics to highlight that they are statistically indistinguishable from GP samples. We also show that the proposed approach is capable of inferring a latent GP observed through a non-linear mapping, and discuss how this can be extended to more complicated non-linear problems.

Preprint. Work in progress.

2 Sequential Gaussian Process Regression

Consider a 1-D (e.g. temporal) zero-mean Gaussian process, $f(t) \sim \mathcal{GP}(0, k(t, t'))$, with covariance function $k(t, t')$. Full batch GP regression involves calculating the covariances between input values within and between some observed set and some test set, and involves the inversion of a potentially large covariance matrix. However, the interpretation of the regression problem as a stochastic differential equation allows for solution in a sequential manner (Hartikainen and Särkkä, 2010).

A Gaussian process can be described as a summation of differential operators driven by a stochastic white-noise process (Särkkä and Solin, 2019), in particular an Itô SDE in companion form: $\mathrm{d}\boldsymbol{f}(t) = \mathbf{F}\boldsymbol{f}(t)\mathrm{d}t + \mathbf{L}\mathrm{d}\boldsymbol{\beta}(t)$. The augmented state is the latent function and its derivatives,

$$\boldsymbol{f}(t) = \begin{bmatrix} f(t) & \frac{\mathrm{d}}{\mathrm{d}t}f(t) & \cdots & \frac{\mathrm{d}^{m-1}}{\mathrm{d}t^{m-1}}f(t) \end{bmatrix}^{\mathrm{T}}.$$

The SDE is driven by a white-noise process, which is itself a Gaussian process by virtue of it being the derivative of Brownian motion so defined with covariance $k_w(t, t') = q\delta_{t,t'}$.

The nature of the covariance function of the GP prior dictates the form of $\mathbf{F}$ and $\mathbf{L}$, the structure of which is linked to the dimension of the latent state $\boldsymbol{f}$ and of the white noise process. For example, the family of Matérn covariance functions, particularly those with half-integer smoothness $\nu = p + 1/2$ are of interest. In the latter case, the covariance functions are differentiable p times, restricting the dimension of $\mathbf{F}$ to $p \times p$. Likewise, the spectral density, q of the white noise process is a scalar and so $\mathbf{L}$ is $p \times 1$.

For the regression problem, with latent function can be mapped as $f(t) = [1\ 0\ \ldots\ 0]\boldsymbol{f}(t)$, and the discretised system is a state-space model and can be solved in linear time with the Kalman filter and Rauch-Tung-Streibel smoother (Hartikainen and Särkkä, 2010).

3 Variational Bridges for Gaussian Process Regression

Consider an Itô process, driven by the SDE $\mathrm{d}\boldsymbol{f}(t) = \boldsymbol{g}(\boldsymbol{f}(t), \theta)\mathrm{d}t + \boldsymbol{c}(\boldsymbol{f}(t), \theta)\mathrm{d}\beta$, where $\boldsymbol{g}$ denotes the drift term and $\boldsymbol{c}^2(\cdot) \triangleq \boldsymbol{c}(\cdot)\boldsymbol{c}(\cdot)^{\mathrm{T}}$ is the diffusion matrix of the SDE. Solutions to the SDE may be intractible and must be approximated in some way, most commonly with iterative solvers. One example is the Euler-Maruyama method, which describes *transition densities* as:

$$f_{k+1} \mid f_k, \theta \sim \mathcal{N}\big(f_k + \Delta_t \boldsymbol{g}(f_k, \theta), \boldsymbol{c}^2(f_k, \theta)\Delta_t\big), \tag{1}$$

where $f_k \triangleq \boldsymbol{f}(t_k)$, the evaluation of $\boldsymbol{f}(\cdot)$ at some discrete time-step t_k; and Δ_t is the step interval, $t_k - t_{k-1}$. The GP model is defined such that $\boldsymbol{g} \triangleq \mathbf{F}_\theta \boldsymbol{f}(t)$; $\boldsymbol{c} \triangleq \mathbf{L}\sqrt{q}$; and θ is the set of hyperparameters of the covariance function. Given some observations $\mathbf{y}$ at times $[\tau_j]_{j=0}^N$, where the observation model is defined $y_j = h(\boldsymbol{f}(\tau_j)) + \varepsilon_j$, conditioning the SDE on the observations gives a Brownian bridge construct with diffusion defined by the GP prior.

In Ryder et al. (2018), the authors use black-box variational approach to designing Brownian bridge constructs. We utilise this approach on the SDE representation of the Gaussian process as described in Section 2, conditioning the diffusion process on observations. The variational approximation of this path, $q_\phi(\boldsymbol{f} \mid \theta)$, has the form

$$q_\phi(\boldsymbol{f} \mid \theta) = \prod_{k=0}^{T} \mathcal{N}\big(f_{k+1} - f_k \mid \tilde{\boldsymbol{g}}(f_k, \mathbf{y}, \theta, t_k, \phi)\Delta_t, \tilde{\boldsymbol{c}}^2(f_k, \mathbf{y}, \theta, t_k, \phi)\Delta_t\big), \tag{2}$$

where $\tilde{\boldsymbol{g}}$ and $\tilde{\boldsymbol{c}}^2$ are the drift and diffusion terms of the variational path.

As an iterative model, driven by Gaussian noise, the variational distribution can also be represented as an implicit generative function $\boldsymbol{f} = \mathfrak{g}(\epsilon, \theta, \phi)$, the result of iteratively solving the reparameterised terms in the product of (2).

The evidence lower bound (ELBO) of the variational parameters, $\mathcal{L}(\phi)$ is maximised using n Monte Carlo realisations of the variational paths and hyperparameters:

$$\hat{\mathcal{L}}(\phi) = \frac{1}{n}\sum_{i=1}^{n} \log \frac{p(\theta^{(i)})p(\boldsymbol{f}^{(i)} \mid \theta^{(i)})p(\mathbf{y} \mid \boldsymbol{f}^{(i)}, \theta^{(i)})}{q_\phi(\theta^{(i)})q_\phi(\boldsymbol{f}^{(i)} \mid \theta^{(i)})}, \tag{3}$$

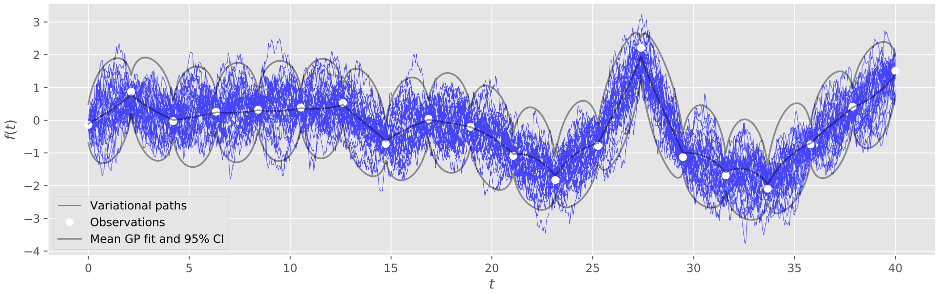

Figure 1: Variational paths after training for 25k epochs in comparison with a full GP regression

where $\boldsymbol{f}^{(i)} = \mathfrak{g}(\epsilon, \theta^{(i)}, \phi_f)$, the implicit model parameterised by a recurrent neural network (RNN).

The variational distribution of the kernel hyperparameters used in the experiments is a mean-field Gaussian approximation, where $q_\phi(\theta) = \prod \mathcal{N}(\theta | \mu_i, s_i)$. The variational parameters are thus the mean and covariance of the Gaussians in the product term: $\phi_\theta = \{(\mu_i, s_i) \,|\, i = 1, \ldots, n_\theta\}$. Likewise, the variational distribution of the latent function $\boldsymbol{f}$ uses an Euler-Maruyama discretisation and constructs Brownian bridges conditioned on the observations, represented by an RNN. The variational parameters ϕ_f in this case are the weights of the RNN.

The generative function parameterised with an RNN represents the terms $\tilde{\boldsymbol{g}}$ and $\tilde{\boldsymbol{c}}$, similar to Ryder et al. (2018); the implementation of the RNN cells uses the same set up and assumptions. The inference is then performed by maximising the ELBO given the posterior of the SDE as defined using the Euler-Maruyama discretisation in (1).

4 Experiments

4.1 GP Regression with Exponential Covariance

In this example, Gaussian process regression is performed with a zero-mean prior and the exponential covariance function: $k(t, t') = \sigma_k^2 \exp(-\lambda |t - t'|)$. A random path is generated, with discrete observations sampled at uniform intervals with additive Gaussian noise with variance $\sigma_y^2 = 0.1$.

The underlying white-noise driven SDE of the GP with an exponential covariance is defined as

$$\mathrm{d}f(t) = -\lambda f(t) + \mathrm{d}\beta(t),$$

with the spectral density of the white-noise process, $q = 2\sigma_k^2\lambda$. The hyperparameters, $\theta = \{\lambda, \sigma_k^2\}$ were approximated with a mean-field Gaussian variational distribution, and updated iteratively with the RNN weights during training.

The 30 variational paths generated from 25,000 epochs of training are shown in Figure 1, along with the sample moments that approximate our variational GP. Overlaid are the moments of a full GP regression using the exponential kernel. We observe visually that mean and credible interval of the GP maps represents a similar fit to the paths.

4.2 Model Criticism

To indicate the reliability of the variational approximation of a Gaussian process, we compare the variational paths generated with samples from the full GP in Figure 1. We map the corresponding paths and samples into a reproducing kernel Hilbert space (RKHS), using a Gaussian kernel, and apply a two-sample test using maximum mean discrepancy (MMD) to provide a metric of the similarity of the distributions generated by the paths and from which the samples are drawn (Gretton et al., 2012).

In Table 1, the MMD^2 value of the variational paths and samples from the full GP fit are shown against the training epoch. As the training time increases, we observe the discrepancy between variational approximation and GP fit decreases, and after 25,000 epochs, the variational paths are indistinguishable from samples from the GP. The results also demonstrate that the similarity between variational paths and GP samples is present even with a small number of observations, due to the conditioned Brownian bridge encoding the GP prior.

Table 1: Unbiased MMD^2 estimations comparing variational paths with samples from a GP for different training epochs and number of observations. Also shown are the respective thresholds indicating 95% rejection

Epoch	10	100	500	1 000	2 500	25 000	Threshold
$T = 6$	0.11106	0.12670	0.05960	0.04843	0.05559	–	0.03713
$T = 20$	0.27305	0.11473	0.06539	0.06962	0.04714	**0.03157**	0.03368

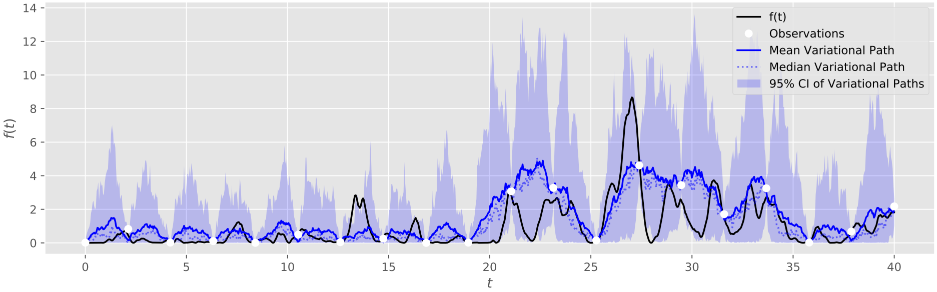

Figure 2: Summary statistics of variational paths with GP prior and non-linear likelihood

4.3 Non-Linear Likelihood

To demonstrate the efficacy in cases where straightforward batch or sequential GP regression is intractible, we perform a similar regression problem on a latent function for which the Gaussian process is related non-linearly.

Consider a latent function with GP prior, u, noisly observed through a non-linear mapping in f:

$$f(t) = u^2(t), \qquad y_k = f(\tau_k) + \varepsilon_k.$$

It's evident here that exact Gaussian process regression on $\mathbf{y}$ is not easily performed directly. In this experiment, we fit variational bridges to $u(t)$ using the approach described in this paper. Again, the exponential covariance was assumed in the prior over our GP in the regression problem.

The mean and 95% confidence interval of the variational paths approximating $f(t)$ are plotted in Figure 2, demonstrating the efficacy of the variational approximation for regression over non-linear observations. We can see this in particular, where the confidence interval captures the trough between the observations at $t \approx 21$ and $t \approx 23$, as well as the peak at $t \approx 27$.

5 Discussion

We demonstrate that we can approximate GP regression using variational bridge constructs by encoding the GP assumptions of a signal as an SDE. The generated variational paths are shown to be indistinguishable from a GP obtained with full batch GP regression, indicating that the approach can produce valid GP samples. While the approach is not intended as a replacement to batch GP regression directly, we show that it is an effective tool for approximating functions where the likelihoods make the regression intractible.

The approach can be futher extended to handle non-linear problems for which there are GP priors involved, utilising the advantages of the variational bridges to solve non-linear SDEs. For example non-linear latent force models (Alvarez et al., 2013), including those problems with non-Gaussian likelihoods, could be approximated with little adjustment to the underlying model. The approach can also make use of past work on using priors with other covariance functions with sequential GP inference, such as periodic and RBF covariances (Solin and Särkkä, 2013; Hartikainen and Särkkä, 2010).

Acknowledgement The authors would like to thank Dennis Prangle and Dougal Sutherland for their input and advice. WW and MAA have been financed by the Engineering and Physical Research Council (EPSRC) Research Project EP/N014162/1.

References

M. A. Alvarez, D. Luengo, and N. D. Lawrence. Linear latent force models using Gaussian processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(11):2693–2705, 2013.

A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773, 2012.

J. Hartikainen and S. Särkkä. Kalman filtering and smoothing solutions to temporal Gaussian process regression models. In *Machine Learning for Signal Processing (MLSP), 2010 IEEE International Workshop on*, pages 379–384. IEEE, 2010.

J. Hartikainen and S. Särkkä. Sequential inference for latent force models. In *Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence*, pages 311–318. AUAI Press, 2011.

J. Hartikainen, M. Seppänen, and S. Särkkä. State-space inference for non-linear latent force models with application to satellite orbit prediction. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pages 723–730, 2012.

J. Hensman, N. Fusi, and N. D. Lawrence. Gaussian processes for big data. In *Uncertainty in Artificial Intelligence*, page 282, 2013.

J. Hensman, A. G. Matthews, and Z. Ghahramani. Scalable variational Gaussian process classification. *Proceedings of Machine Learning Research*, 38:351–360, 2015.

H. Nickisch, A. Solin, and A. Grigorievskiy. State space Gaussian processes with non-Gaussian likelihood. *arXiv preprint arXiv:1802.04846*, 2018.

T. Ryder, A. Golightly, A. S. McGough, and D. Prangle. Black-box variational inference for stochastic differential equations. *arXiv preprint arXiv:1802.03335*, 2018.

S. Särkkä and A. Solin. *Applied Stochastic Differential Equations*. Institute of Mathematical Statistics Textbooks. Cambridge University Press, 2019.

A. Solin and S. Särkkä. Infinite-dimensional Bayesian filtering for detection of quasiperiodic phenomena in spatiotemporal data. *Physical Review E*, 88(5):052909, 2013.

S. Särkkä. *Bayesian Filtering and Smoothing*. Institute of Mathematical Statistics Textbooks. Cambridge University Press, 2013.