Automated Augmented Conjugate Inference for Non-conjugate Gaussian Process Models

Théo Galy-Fajou
Technical University of Berlin

Florian Wenzel
Google Research*

Manfred Opper
Technical University of Berlin

Abstract

We propose *automated augmented conjugate inference*, a new inference method for non-conjugate Gaussian processes (GP) models. Our method automatically constructs an auxiliary variable augmentation that renders the GP model conditionally conjugate. Building on the conjugate structure of the augmented model, we develop two inference methods. First, a fast and scalable stochastic variational inference method that uses efficient block coordinate ascent updates, which are computed in closed form. Second, an asymptotically correct Gibbs sampler that is useful for small datasets. Our experiments show that our method are up two orders of magnitude faster and more robust than existing state-of-the-art black-box methods.

1 INTRODUCTION

Developing automated yet efficient Bayesian inference methods for Gaussian process (GP) models is a challenging problem that has attracted considerable attention within the probabilisitic machine learning community (Salimbeni et al., 2018; Wenzel et al., 2019). A GP defines a distribution over functions and can be used as a flexible building block to develop expressive probabilistic models. By choosing an appropriate likelihood function on top of a latent GP, a variety of interesting models is obtained, which are successfully used in several application areas including robotics (Beckers et al., 2019), facial behavior analysis (Eleftheriadis et al., 2017) and electrical engineering (Pandit and Infield, 2018). For instance, using a logistic likelihood leads to a binary GP classification model, and using a Student-t likelihood can be used for robust regression.

The main challenge in these models is to infer the latent GP given a general non-Gaussian likelihood. Methods that are more generally applicable often treat the model as a black box and are based on sampling or numerical quadrature, thus, preventing efficient optimization (Hensman et al., 2015; Salimbeni et al., 2018). On the other side. a lot of methods focus on special cases of GP models (i.e. special likelihood functions) by exploiting model specific properties, e.g. binary classification (Polson et al., 2013).

In this work, we develop *automated augmented conjugate inference (*AACI*)*. AACI is an efficient inference framework, which is applicable to a large class of GP models that use a super-Gaussian likelihood[1]. It automatically exploits specific properties of the likelihood leading to an inference algorithm that is up to two orders of magnitudes faster than the state of the art.

Our approach builds on an auxiliary variable augmentation of the model: we add a latent variable to the model such that the original model is recovered when this variable is integrated out. We consider an augmentation that renders the model conditionally conjugate. In a conditionally conjugate model, all complete conditional distributions (the posterior distribution of one random variable given all the others), can be computed in closed form. Moreover, we show that inference in the augmented conditionally conjugate model is much easier than in the original model and demonstrate superior performance over the state of the art.

Building on the conditionally conjugate augmentation, AACI provides two options for inference: a scalable variational inference method based on efficient closed-form coordinate ascent updates and an exact Gibbs sampling method, which is useful on smaller datasets.

Our main contributions are as follows:

- We introduce AACI: an automated inference method for GP models with a super-Gaussian likelihood.
- We propose two inference modules: augmented variational inference, which scales to large datasets contain-

---

*Work done while at TU Berlin

[1]The definition of the family of super-Gaussian likelihoods is given in Section 3.
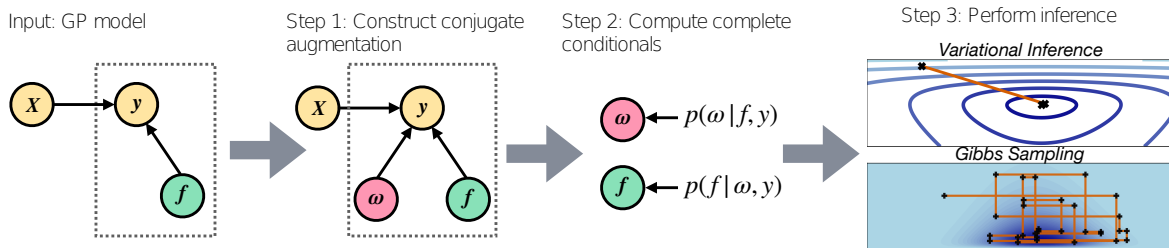
*Figure 1. Automated augmented conjugate inference (*AACI*)* performs automated efficient inference in non-conjugate Gaussian process models. In the first step, AACI translates the GP model into an augmented model that is conditionally conjugate. In the second step, the complete conditionals are computed in closed form. In the final step, AACI provides two options: (A) fast stochastic variational inference based on coordinate ascent updates, which easily scales to big datasets and (B) an asymptotically exact Gibbs sampler, which provides high quality samples from the true posterior but is limited to smaller datasets.

ing millions of instances and an exact Gibbs sampler, which is useful for small datasets.

- The experiments demonstrate that the augmented variational inference module of AACI outperforms the state of the art in terms of speed by up to two orders of magnitude while being competitive in terms of prediction performance. The Gibbs sampler module leads to a much better efficient sample size while still being up to ten times faster than Hamiltonian Monte Carlo.

The paper is structured as follows: Section 2 gives a high-level overview about our novel inference method AACI. In Section 3, we provide a detailed discussion of the algorithm and proof that our approach indeed leads to conditionally conjugate models. We discuss related work in Section 4 and show our experimental results in Section 5. Finally, Section 6 concludes and lays out future research directions. Our source code for the experiments is included in a gitgub repository[2].

2 AUTOMATED AUGMENTED CONJUGATE INFERENCE

Let $X = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^\top \in \mathbb{R}^{n \times d}$ be a matrix of data points and $\boldsymbol{y} = (y_1, \ldots, y_n) \in \mathbb{R}^n$ the corresponding target values. The goal is to learn a mapping from the input points to the target values via a latent function f. We assume a prior GP distribution (with mean prior $\boldsymbol{\mu}_0$ and covariance function $k(x, x')$) on the latent function and the data labels $\boldsymbol{y} = (y_1, \ldots, y_n)$ are connected to f via a factorizable likelihood

$$p(f) = \mathrm{GP}(f|\boldsymbol{\mu}_0, k), \quad p(\boldsymbol{y}|f, X) = \prod_{i=1}^{n} p(y_i|f(\boldsymbol{x}_i)).$$

The key inference challenge in the GP models is to compute the posterior distribution of the latent function

$$p(f|y) = \frac{p(y|f)p(f)}{\int p(y|f)p(f)\mathrm{d}y}.$$

This is a challenging problem. Inference in GP models scale cubically in the number of data points and is intractable for non-Gaussian likelihoods.

Ideally, we would like an efficient inference method that is not hand-tailored to a specific type of likelihood and hence allows for experimenting with different types of GP models on big datasets in a scalable manner. Thus, we need a flexible inference method that works for a large class of likelihoods, is fast and ideally does not involve inefficient black box approaches as approximating the objective by sampling.

2.1 Automated Augmented Conjugate Inference

We introduce the *automated augmented conjugate inference (*AACI*)* to achieve this goal. AACI accelerates training of GP models whose likelihood is in the family of super-Gaussian likelihood functions.

AACI translates the intractable non-conjugate model into an easier, conditionally conjugate model by adding auxiliary random variables to the model. Inference in conditionally conjugate models is a classic and well-studied problem (Bishop, 2006). Because of the special structure of conditionally conjugate models, many efficient inference methods exist (Wang and Blei, 2013). Based on the automatically constructed augmentation, we propose an efficient variational inference method using coordinate ascent updates and a Gibbs sampler.

The inference pipeline of **AACI.** AACI consists of three steps. In the first step, a conjugate augmentation of the model is constructed by adding auxiliary variables ω to the

---

[2]https://github.com/theogf/AutoConjGP_Exp

model. Then, the complete conditional distributions of the latent function f and auxiliary variables ω are computed. In the final step, we provide two options to perform inference.

The *variational inference* (VI) module of AACI performs block coordinate ascent updates, computed in closed form. The updates are much more efficient than ordinary Euclidean gradient updates, which are used in most previous approaches. The *Gibbs sampling* module of AACI builds on the complete conditional distributions and provides exact samples from the true posterior. For each type of likelihood, the sampler is automatically constructed.

The inference pipeline of AACI is summarized in Fig. 1. In the following, we give an overview of how each module of our inference pipeline works and provide the details in Section 3.

(1) Augmenting the model. The first step of our inference framework constructs an *auxiliary variable augmentation* that renders the model *conditionally conjugate*. Our augmentation approach finds a Gaussian scale mixture representation of the intractable likelihood

$$p(y_i|f_i) = \int p(y_i|f_i, \omega_i)p(\omega_i)\mathrm{d}\omega, \qquad (1)$$

where $p(y_i|f_i, \omega_i)$ is an unnormalized Gaussian distribution in f_i with precision ω_i and $p(\omega_i)$ is the prior distribution of the auxiliary variable. The construction of the distribution $p(\omega)$ is based on an inverse Laplace transformation and is discussed in Section 3.1.

Building on Eq. 1, we augment the GP model by a set of auxiliary variables $\omega = (\omega_1, \dots, \omega_n)$ leading to the augmented joint distribution

$$p(\boldsymbol{y}, \boldsymbol{f}, \boldsymbol{\omega}) = \prod_i p(y_i|f_i, \omega_i)p(\omega_i)p(\boldsymbol{f}), . \qquad (2)$$

The auxiliary variable augmentation is constructed in a way such that the augmented model is *conditionally conjugate*, i.e. the complete conditional distributions $p(\boldsymbol{\omega}|\boldsymbol{f}, \boldsymbol{y})$ and $p(\boldsymbol{f}|\boldsymbol{\omega}, \boldsymbol{y})$ are in the same family as their associated priors.

(2) Computing the complete conditionals. The complete conditionals of $\boldsymbol{f}$ and the auxiliary variables ω_i are computed in closed form and are given by

$$p(\boldsymbol{f}|\boldsymbol{y}, \boldsymbol{\omega}) = \mathcal{N}(\boldsymbol{f}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$$
$$p(\omega_i|f_i, y_i) = \pi_\varphi(\omega_i|c_i),$$

where φ is a function determined by the type of the likelihood (see Eq. 4) and the parameters $\boldsymbol{\mu}, \boldsymbol{\Sigma}, c_i$ have closed-form expressions and are described in Section 3.2. The distribution family $\pi_\varphi(\omega|c)$ is derived by an exponential tilting of the prior distribution $p(\omega)$ and is discussed in Section 3.2.

(3a) Augmented variational inference. In step 3, AACI provides two options to perform inference. We first discuss the variational inference module, which approximates the posterior by a variational distribution and easily scales to big datasets.

We assume a mean-field variational distribution, where the latent GP $\boldsymbol{f}$ and the auxiliary variables $\boldsymbol{\omega}$ are decoupled, i.e. $q(\boldsymbol{f}, \boldsymbol{\omega}) = q(\boldsymbol{f})q(\boldsymbol{\omega})$. The optimal variational distribution of $\boldsymbol{\omega}$ naturally factorizes, i.e. $q(\boldsymbol{\omega}) = \prod_i q(\omega_i)$. Following standard results (Bishop, 2006) the variational distributions can be iteratively optimized by the block-coordinate ascent updates:

$$q(\boldsymbol{f}) \propto \exp\left(\mathbb{E}_{q(\boldsymbol{\omega})}\left[\log p(\boldsymbol{f}|\boldsymbol{\omega}, \boldsymbol{y})\right]\right)$$
$$q(\omega_i) \propto \exp\left(\mathbb{E}_{q(\boldsymbol{f})}\left[\log p(\omega_i|\boldsymbol{f}, \boldsymbol{y})\right]\right). \qquad (3)$$

In Section 3.3, we show that these updates are given in closed form and can be computed efficiently without resorting to numerical methods. To scale to big datasets we employ SVI (Hoffman et al., 2013) and replace the original latent GP f by Titsias (2009) sparse approximation building on inducing points .

(3b) Exact inference via Gibbs sampling. Building on the conditionally conjugate augmentation, it is straightforward to derive a Gibbs sampler. In order to sample from the exact posterior, we alternate between drawing a sample from each complete conditional distribution

$$\boldsymbol{\omega}^t \sim p(\boldsymbol{\omega}|\boldsymbol{f}^{t-1}, \boldsymbol{y}),$$
$$\boldsymbol{f}^t \sim p(\boldsymbol{f}|\boldsymbol{\omega}^t, \boldsymbol{y}).$$

The augmented variables are naturally marginalized out and the latent GP samples $\{\boldsymbol{f}^t\}$ will be from the true posterior $p(\boldsymbol{y}|\boldsymbol{f})$. As we empirically show in Section 5.1, the Gibbs sampler leads to very fast mixing and outperforms standard Hamiltonian Monte Carlo sampling.

3 ALGORITHM DETAILS

Here we provide the details on the *automated augmented conjugate inference (*AACI*)* algorithm. We start by specifying the class of GP models that we consider in our framework. We then discuss the technical details of AACI and proof that the automatically constructed augmentation indeed leads to a conditionally conjugate model.

GP Models with a super-Gaussian likelihood. AACI can be applied to GP models, where the likelihood is within the class of super-Gaussian likelihoods. A super-Gaussian likelihood is of the form

$$p(\boldsymbol{y}|\boldsymbol{f}; \theta) = C(\theta)e^{g(\boldsymbol{y};\theta)^\top \boldsymbol{f}}\varphi(||h(\boldsymbol{f}, \boldsymbol{y})||_2^2), \qquad (4)$$

where θ are hyperparameters of the likelihood, $C(\theta)$ is the normalizing constant, $g(\boldsymbol{y}; \theta)$ is an arbitrary function, φ is

a *positive definite radial (PDR) function*[3], and h is a linear function in $\boldsymbol{f}$, such that we can write

$$||h(\boldsymbol{f},\boldsymbol{y})||_2^2 = \alpha(\boldsymbol{y},\theta) - \beta(\boldsymbol{y},\theta)^\top \boldsymbol{f} + \gamma(\boldsymbol{y},\theta)||\boldsymbol{f}||_2^2, \quad (5)$$

where α, β, γ are arbitrary functions. We omit θ in the later derivations for clarity.

Many interesting models are instances of super-Gaussian likelihood GP models. In Table 1, we present several likelihood functions with their corresponding parameter settings of the super-Gaussian likelihood as given in Eq. 4.

Constructing new likelihoods. Using Eq. 4, we can also construct novel likelihood functions based on existing kernel functions. In this paper we propose the Matern 3/2 likelihood.

3.1 Step 1: Conjugate augmentation

Given the likelihood of the model, AACI constructs a conditionally conjugate auxiliary variable augmentation as follows. We first define a family of distribution $\pi_\varphi(\omega|c)$, which will be useful for constructing the augmentation.

For the case $c = 0$, the distribution $\pi_\varphi(\omega|0)$ is defined by the inverse Laplace transform of $\varphi(\cdot)$,

$$\pi_\varphi(\omega|0) = \mathcal{L}^{-1}\left\{\varphi(\cdot)\right\}(\omega). \quad (6)$$

The inverse Laplace is the inverse mapping of the Laplace transformation and can be computed by the Bromwich integral formula[4] (Debnath and Bhatta, 2014) and it defines a valid density in our setting (see proof of Theorem 1). Remarkably, we will see that for the final updates of our algorithm, we do not need to compute the inverse Laplace transformation explicitly.

We generalize the base distribution $\pi_\varphi(\omega|0)$ by applying an exponential tilting:

$$\pi_\varphi(\omega|c) = \frac{e^{-c^2\omega}\pi_\varphi(\omega|0)}{\varphi(c^2)}, \quad (7)$$

where $c \in \mathbb{R}$.

Theorem 1. *A GP model with a super-Gaussian likelihood (of the form of Eq. 4) is rendered **conditionally conjugate** by the auxiliary variable augmentation $p(\boldsymbol{y},\boldsymbol{f},\boldsymbol{\omega};\theta) = p(\boldsymbol{y}|\boldsymbol{f},\boldsymbol{\omega};\theta)p(\boldsymbol{f})p(\boldsymbol{\omega})$. The augmented likelihood is*

$$p(\boldsymbol{y}|\boldsymbol{f},\boldsymbol{\omega};\theta) = C(\theta)\exp\left(g(\boldsymbol{y};\theta)^\top \boldsymbol{f} - ||h(\boldsymbol{f},\boldsymbol{y})||_2^2\boldsymbol{\omega}\right)$$

---

[3]φ is a positive definite radial function if $\varphi(r)$ is completely monotone for all $r \geq 0$ and $\lim_{r\to 0}\varphi(r) = 1$.

[4]The inverse Laplace transformation of a function $\varphi(\cdot)$ can be computed by $\mathcal{L}^{-1}\left\{\varphi(\cdot)\right\}(\omega) = \lim_{T\to\infty}\frac{1}{2\pi i}\int_{b-iT}^{b+iT}e^{r\omega}\varphi(r)dr$, where b can be arbitrarily chosen but has to be larger than the real part of all singularities of φ.

and the prior distribution of the auxiliary variables is

$$p(\boldsymbol{\omega}) = \pi_\varphi(\boldsymbol{\omega}|0).$$

Proof: We first apply Schoenberg's theorem (Schoenberg, 1938), which states that a function $\mathbb{R}^d \ni \boldsymbol{x} \to \varphi(\|\boldsymbol{x}\|_2^2)$ is a PDR function for any dimension $d > 0$ if and only if $\varphi(r)$ is a completely monotone function on the domain $r \geq 0$.

A completely monotone function $\varphi(\cdot)$ has the property that it is infinitely differentiable and its derivatives have an alternating sign (Bernstein et al., 1929), i.e.

$$(-1)^k\varphi^{(k)}(r) > 0, \quad r \in [0,+\infty), \; k = 0,1,2,\dots. \quad (8)$$

As a direct consequence, $\varphi(\cdot)$ is a positive, decreasing, and convex function and the first derivative of $\varphi(\cdot)$ is a concave function.

Building on these properties, Widder (1946) states that we can rewrite $\varphi(\|h(f,y)\|_2^2)$ as a Gaussian scale-mixture

$$\varphi\left(\|h(f,y)\|_2^2\right) = \int_0^\infty e^{-\|h(f,y)\|_2^2\omega}d\mu(\omega), \quad (9)$$

with respect to a Borel measure $\mu(\omega)$. We apply the monotone convergence theorem (Yeh, 2006), which gives that $\mu(\omega)$ is even a probability measure iff $\lim_{r\to 0}\varphi(r) = 1$. Since we have a probability measure, we write $d\mu(\omega) = p(\omega)d\omega$ and which leads to the equality $\varphi(r) = \mathcal{L}\{p(\omega)\}(r)$, where $\mathcal{L}$ denotes the Laplace transformation. The inverse Laplace transformation gives the density of the auxiliary variable $p(\omega) = \mathcal{L}^{-1}\{\varphi(r)\}(\omega) = \pi_\varphi(\omega|0)$.

Therefore we can rewrite the super-Gaussian likelihood Eq. 4 as :

$$p(y|f) = C(\theta)\int_0^\infty e^{-g(y)f-\|h(f,y)\|_2^2\omega}p(\omega)d\omega. \quad (10)$$

Adding the auxiliary variable ω with prior $p(\omega)$ to the model, we obtain the augmented likelihood $p(\boldsymbol{y}|\boldsymbol{f},\boldsymbol{\omega};\theta) = C(\theta)\exp\left(g(\boldsymbol{y};\theta)^\top \boldsymbol{f} - ||h(\boldsymbol{f},\boldsymbol{y})||_2^2\boldsymbol{\omega}\right)$.

Since the function $g(\boldsymbol{y};\theta)^\top \boldsymbol{f} - ||h(\boldsymbol{f},\boldsymbol{y})||_2^2\boldsymbol{\omega}$ is by definition quadratic in $\boldsymbol{f}$ the augmented likelihood is proportional to an (unnormalized) Gaussian distribution in $\boldsymbol{f}$, hence, conditionally conjugate in $\boldsymbol{f}$.

For the augmented variable ω_i, the likelihood $p(y|\omega,f)$ act as an exponential tilting of $p(\omega)$ and the full conditional in ω will stay in the same family of distributions. QED.

3.2 Step 2: Complete Conditionals

Since the augmented model (Section 3.1) is conditionally conjugate, the complete conditional distribution are in the

Likelihood	Full form	$g(f, y)$	$h(f, y)$	$\varphi(r)$
Student-t	$\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu}\pi\sigma\Gamma\left(\frac{\nu}{2}\right)}\left(1 + \frac{(y-f)^2}{\nu\sigma^2}\right)^{-\frac{\nu+1}{2}}$	0	$\frac{f-y}{\sigma}$	$\left(1 + \frac{r}{\nu}\right)^{-\frac{\nu+1}{2}}$
Laplace	$\frac{1}{2\beta}\exp\left(-\frac{|y-f|}{\beta}\right)$	0	$f - y$	$\exp\left(-\frac{\sqrt{r}}{\beta}\right)$
Logistic	$\frac{1}{2}\exp\left(\frac{yf}{2}\right)\cosh^{-1}\left(\frac{|yf|}{2}\right)$	$\frac{yf}{2}$	$\frac{f}{2}$	$\cosh^{-1}\left(\sqrt{r}\right)$
Bayesian SVM	$\exp\left((yf - 1) - |1 - yf|\right)$	yf	$1 - yf$	$\exp(-\sqrt{r})$
Matern 3/2	$\frac{\sqrt{3}}{4\rho}(1 + \frac{\sqrt{3}|y-f|}{\rho})\exp(-\frac{\sqrt{3}|y-f|}{\rho})$	0	$f - y$	$(1 + \frac{\sqrt{3r}}{\rho})\exp(-\frac{\sqrt{3r}}{\rho})$

Table 1. Many interesting GP models are members of the super-Gaussian likelihood family introduced in Section 3. We display the full likelihood and the corresponding terms of the super-Gaussian likelihood as described in Eq. 4. Some models were already considered independently but our approach provides a unified view.

same family as their associated prior distributions and are given in closed form.

Theorem 2. *The **complete conditional distributions** of the augmented model presented in Section 3.1 are given by*

$$p(\omega_i|f_i, y_i) = \pi_\varphi\left(\omega_i | \|h(f_i, y_i)\|_2\right),$$
$$p(\boldsymbol{f}|\boldsymbol{y}, \boldsymbol{\omega}) = \mathcal{N}\left(\boldsymbol{f}|\boldsymbol{\mu}, \boldsymbol{\Sigma}\right), \tag{11}$$

where $\boldsymbol{\Sigma} = \left(\mathrm{diag}\left(2\boldsymbol{\omega} \circ \gamma(\boldsymbol{y})\right) + K^{-1}\right)^{-1}$ *and* $\boldsymbol{\mu} = \boldsymbol{\Sigma}\left(g(\boldsymbol{y}) + \boldsymbol{\omega} \circ \beta(\boldsymbol{y}) + K^{-1}\boldsymbol{\mu}_0\right)$, $\circ$ *denotes the Hadamard product and the function* $h(\cdot)$ *is given by the form of likelihood (see Eq.5).*

The proof is given in Appendix A.1

3.3 Step 3: Efficient inference

In the final step of our inference pipeline, we leverage the conditionally conjugate structure of the augmented model and derive two inference methods. First, we propose a scalable stochastic variational inference (SVI) method that builds on efficient block coordinate ascent updates (CAVI) updates, computed in closed form. Second, we develop a Gibbs sampling scheme that generates samples from the exact posterior.

3.3.1 Augmented variational inference

We implement the classic stochastic variational inference (SVI) algorithm for conditionally conjugate models described by Hoffman et al. (2013), which builds on block coordinate ascent updates. The updates can be interpreted as natural gradient updates and are much more efficient than ordinary Euclidean gradient updates (Amari, 1998).

Variational approximation. We approximate the posterior distribution of the latent GP values by assuming a decoupling between $\boldsymbol{f}$ and $\boldsymbol{\omega}$. The family of the optimal variational distribution can be easily determined by averaging the complete conditionals in log-space, as given in Eq. 3 (see

e.g. Blei et al., 2017). From the above decoupling assumption, it follows that the optimal variational posterior is in the variational family

$$q(\boldsymbol{f}, \boldsymbol{\omega}) = q(\boldsymbol{f})\prod_{i=1}^{N} q(\omega_i), \tag{12}$$

where $q(\boldsymbol{f}) = \mathcal{N}\left(\boldsymbol{f}|\boldsymbol{m}, \boldsymbol{S}\right)$ and $q(\omega_i) = \pi_\varphi(\omega_i|c_i)$ and $\boldsymbol{m}, \boldsymbol{S}$ and $\boldsymbol{c}$ are the *variational parameters*.

Variational updates. We start with deriving the variational updates for the variational Gaussian distribution,

$$q(\boldsymbol{f}) \propto \exp\left[\mathbb{E}_{q(\boldsymbol{\omega})}\left[\log p(\boldsymbol{f}|\boldsymbol{\omega}, \boldsymbol{y})\right]\right]$$
$$\propto \exp\left[\sum_i g(y_i)f_i - \|h(f_i, y_i)\|_2^2\mathbb{E}_{q(\omega_i)}\left[\omega_i\right]\right]p(\boldsymbol{f})$$

Computing the variational updates of $q(\boldsymbol{f})$ boils down to computing the first moment of $\boldsymbol{\omega}$. Remarkably, the moments of π_φ can be computed without computing the closed-form density of π_φ explicitly, i.e. without evaluating the inverse Laplace transformation of φ (Eq. 6).

The moments can be computed by differentiating the moment generating function, which is itself a Laplace transform. For our algorithm, we only need the first moment of ω, which is given by

$$\mathbb{E}_{q(\omega)}\left[\omega\right] = \left.\frac{d\mathcal{L}\left\{q(\omega)\right\}(-t)}{dt}\right|_{t=0} = -\frac{\varphi'(c^2)}{\varphi(c^2)} = \overline{\omega},$$

which can be cheaply computed via automatic differentiation.

The updates for the variational distribution of the auxiliary variables $q(\boldsymbol{\omega})$ are computed as follows.

$$q(\omega_i) \propto \exp\left[-\mathbb{E}_{q(f_i)}\left[\|h(f_i, y_i)\|_2^2\right]\omega_i + \log p(\omega_i)\right]$$
$$\propto \exp\left(-\mathbb{E}_{q(f_i)}\left[\|h(f_i, y_i)\|_2^2\right]\omega_i\right)p(\omega_i)$$
$$= \pi_\varphi(\omega_i|\sqrt{\mathbb{E}_{q(f_i)}\left[h(f_i, y_i)^2\right]}).$$

We get then the update $c_i = \sqrt{\mathbb{E}_{q(f_i)}\left[\|h(f_i, y_i)\|_2^2\right]}$, which can be easily computed in closed form since $\|h(f_i, y_i)\|_2^2$ is a quadratic function of f_i.

The coordinate ascent variational inference (CAVI) method is summarized in Algorithm 1.

Algorithm 1 Augmented Variational Inference

Input: Data $(\mathbf{X}, \boldsymbol{y})$, GP model $p(\boldsymbol{y}|\boldsymbol{f})$, kernel k
Output: Approximate posterior $q(f) = \mathcal{N}(f \mid \boldsymbol{m}, \boldsymbol{S})$
for iteration $t = 1, 2, \ldots,$ **do**
  # Local updates:
  for $i \in 1 : N$ **do**
    $c_i = \sqrt{\mathbb{E}_{q(f)}\left[h(f_i, y_i)^2\right]}$
    $\overline{\omega}_i = \mathbb{E}_{q(\omega_i)}\left[\omega_i\right] = -\varphi'(c_i^2)/\varphi(c_i^2)$
  end for
  # Coordinate ascent updates (CAVI):
  $\boldsymbol{S} \leftarrow \left(\mathrm{diag}\left(2\overline{\boldsymbol{\omega}} \circ \gamma(\boldsymbol{y})\right) + K^{-1}\right)^{-1}$
  $\boldsymbol{m} \leftarrow \boldsymbol{S}\left(K^{-1}\mu_0 + g(\boldsymbol{y}) + \overline{\boldsymbol{\omega}} \circ \beta(\boldsymbol{y})\right)$
end for

Sparse GP approximation. To scale our method to big datasets, we approximate the latent GP $\boldsymbol{f}$ by a *sparse Gaussian process* building on *inducing points*. We introduce M inducing points $\boldsymbol{u}$ and connect the GP values with the inducing points via the joint prior distribution $p(\boldsymbol{f}, \boldsymbol{u})$ given in Titsias (2009). The introduction of inducing points preserves conditional conjugacy and allows for mini-batch sampling of the data (stochastic variational inference). This scales the algorithm to big datasets and has the computational complexity $\mathcal{O}(M^3)$. The SVI version of our algorithm only slightly changes the updates that are presented in Algorithm 1. It is deferred to Appendix A.3.

3.3.2 Gibbs sampling

To sample from the exact posterior distribution, a Gibbs sampling scheme alternates between sampling from the complete conditional distributions. In the following we propose a sampling scheme for the distribution family $\pi_\varphi(\omega|c)$ that is automatically constructed given the PDR function of the likelihood $\varphi(\cdot)$

The distribution class π_φ is defined in Eq. 6 and is based on the inverse Laplace transform of $\varphi(\cdot)$. However there is no general approach to compute the inverse Laplace in closed form (Cohen, 2007). We circumvent this issue by proposing an algorithm that only evaluates the inverse Laplace transformation point-wise but does not need access to its full analytical form. We apply the method proposed by Ridout (2009), which build on the fact that the cumulative density function (CDF) $F_{\pi_\varphi(\omega|c)}(\cdot)$ can be computed via the inverse Laplace transform of a scaled (forward) Laplace trans-

form,

$$F_{\pi_\varphi(\omega|c)}(x) = \mathcal{L}^{-1}\left\{\frac{\mathcal{L}\left\{\pi_\varphi(\omega|c)\right\}(s)}{s}\right\}(x)$$
$$= \mathcal{L}^{-1}\left\{\frac{\varphi(s+c^2)}{s\varphi(c^2)}\right\}(x).$$

To generate samples from $\pi_\varphi(\omega|c)$, we first generate a uniform sample $u \sim \mathcal{U}[0,1]$ and then push it through the inverse CDF, $\omega = F_{\pi_\varphi(\omega|c)}^{-1}(u)$ (Devroye, 1986) Finally, to compute the inverse CDF, we solve a fixed point problem using the modified Newton-Raphson method described by Ridout (2009). We solve the equation $F_{\varphi(c)}(\omega) = u$ by repeatedly setting $\omega \leftarrow \omega - F_{\varphi(c)}(\omega)/\pi_\varphi(\omega|c)$ until reaching convergence. We numerically approximate the (forward) CDF $F_{\varphi(c)}(\omega)$ by the cheap trapezoidal method introduced in Abate et al. (2000), which has error guarantees. The cost of this process is negligible against the matrix inversion for sampling $\boldsymbol{f}$. All steps are summarized in Algorithm 2.

Note that for some likelihood functions (e.g. the logistic likelihood function), the inverse Laplace transform can be derived analytically and the steps described above can be optimized by using an existing the sampler for the corresponding complete conditional distribution.

Algorithm 2 Gibbs Sampling

Input: Data $(\mathbf{X}, \boldsymbol{y})$, GP model $p(\boldsymbol{y}|\boldsymbol{f})$, kernel k
Output: Posterior samples $\{\boldsymbol{f}^t\} \sim p(\boldsymbol{f} \mid \boldsymbol{y})$
for sample index $t = 1, 2, \ldots,$ **do**
  # Sample $\boldsymbol{\omega} \sim p(\boldsymbol{\omega}|\boldsymbol{f}, \boldsymbol{y})$:
  for $i \in 1 : N$ **do**
    Compute $c_i = \|h(f_i, y_i)\|_2$
    Sample $u_i \sim \mathcal{U}[0,1]$
    # Compute inverse cdf $\omega_i = F_{\pi_\varphi(c_i)}^{-1}(u_i)$:
    Initialize $\omega_i > 0$
    while $|\widetilde{F}_{\pi_\varphi(c_i)}(\omega_i) - u_i| > \epsilon$ **do**
      Approximate $\widetilde{F}_{\pi_\varphi}(\omega_i), \widetilde{\pi}_\varphi(\omega_i|c_i)$(see Sec.3.3.2)
      $\omega_i \leftarrow \omega_i - \frac{\widetilde{F}_{\pi_\varphi(c_i)}(\omega_i)}{\widetilde{\pi}_\varphi(\omega_i|c_i)}$
    end while
  end for
  # Sample $\boldsymbol{f} \sim p(\boldsymbol{f}|\boldsymbol{\omega}, \boldsymbol{y})$:
  $\boldsymbol{\Sigma} = \left(\mathrm{diag}\left(2\boldsymbol{\omega} \circ \gamma(\boldsymbol{y})\right) + K^{-1}\right)^{-1}$
  $\boldsymbol{\mu} = \boldsymbol{\Sigma}\left(K^{-1}\mu_0 + g(\boldsymbol{y}) + \boldsymbol{\omega} \circ \beta(\boldsymbol{y})\right)$
  Sample $\boldsymbol{f}^t \sim \mathcal{N}\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$
end for

---

4 RELATED WORK

Inference for non-conjugate likelihoods is not a new topic and there have been many works to deal efficiently with the problem.

Scale mixtures of normals. The Gaussian scale-mixture formulation is well known in statistics and have been explored more recently by Gneiting (1997, 1999). Palmer (2006); Palmer et al. (2006) started to generalize it for a machine learning use but did not explore the probability side of the augmentation.

Black-box variational inference. One of the most popular approach for variational inference in the recent years is to optimize the ELBO for an arbitrary model by computing gradients estimates via sampling or quadrature, e.g. Salimbeni et al. (2018); Mohamed et al. (2019). However these methods do not exploit the structure of the model and can be less efficient.

Sampling methods. Sampling is not a popular method for GP models since f is high-dimensional and the posterior is usually highly correlated (Lawrence et al., 2009). But as for many Bayesian models, Hamiltonian Monte Carlo is a good candidate (Titsias et al., 2008).

Likelihood approximation. Jaakkola and Jordan (2000) propose a variational approach purely based on optimization, using the partial convexity of the likelihood. Our method recovers their results, but coming from a probabilistic perspective. We show in Appendix A.5, the equivalence with their approach. Khan and Lin (2017) exploit existing partial conjugacy in the model and rely on the assumption that part of the joint posterior can be rewritten as an exponential family. Their approach is complementary to ours and could be combined for solving more complex models.

Use cases of the augmented model. Different applications of the augmentation technique for specific likelihoods have been explored in multiple papers: Jylänki et al. (2011) applied the augmentation on the Student-t likelihood with Gaussian Processes. Polson et al. (2013) developed an approach with the logistic likelihood, this work was further expanded by Wenzel et al. (2019) to big data. The augmentation done on the Bayesian Support Vector Machine of Polson et al. (2011) and scaled up by Wenzel et al. (2017), is similar to our method but is based on a different augmentation approach. Note that our method covers all these cases exactly but do not rely on any manual derivations.

5 EXPERIMENTS

In this section we answer the following questions empirically:

- How does the Gibbs sampling scheme compare to other sampling methods?
- What is lost in variational inference by approximating an additional variable?
- And what is the gain in speed?

We explore four different cases. We use three regression models with different likelihood functions: a Laplace like-

Likelihood/Method		MH	HMC	**Gibbs**
Logistic	Time/Sample (s)	**0.001**	0.041	0.01
	Lag 1	0.996	0.53	**0.11**
	Gelman	1.38	**1.00**	**1.00**
Student-t	Time/Sample (s)	**0.003**	0.573	0.028
	Lag 1	1.0	0.857	**0.04**
	Gelman	1.51	1.00	**1.00**
Laplace	Time/Sample (s)	**0.002**	0.082	0.028
	Lag 1	0.995	0.931	**0.26**
	Gelman	1.44	1.01	**1.00**
Matern 3/2	Time/Sample (s)	**0.005**	0.15	0.029
	Lag 1	0.997	0.995	**0.05**
	Gelman	1.59	1.10	**1.00**

Table 2. Sampling time and diagnostics of Gibbs Sampling, naive Metropolis-Hastings and Hamiltonian Monte-Carlo. The Gelman test indicates the inter-chain correlation and should be close to 1.

lihood, a Student-t likelihood, a new likelihood inspired by the Matern 3/2 kernel (Rasmussen, 2003) and one classification model with a logistic likelihood. All the mathematical details of these augmentations are deferred to the Appendix A.6. For the two first experiments we use a full GP without inducing points to have a cleaner analysis of the effect of the augmentation. For all experiments we use a squared exponential kernel with automatic relevance determination: $k(x, x') = \exp(-\sum_{d=1}^{D}(x_d - x'_d)^2/\theta_d^2)$. For the two first experiments we use datasets from the UCI repository (Dua and Graff, 2017) : the Boston housing dataset ($N = 506, D = 14$) for regression and the Heart dataset ($N = 303, D = 14$) for classification. For the last experiment we use the Protein dataset ($N = 45730, D = 9$) and the Airline dataset ($N = 190K, D = 7$) for regression and the Covtype dataset ($N = 581K, D = 54$) and the SUSY dataset ($N = 5M, D = 18$) for classification. We normalize the input features to mean 0 and variance 1.

5.1 Gibbs sampling mixing

Our approach leads to a Gibbs sampling algorithm that provides samples from the true posterior of the original model. We compare our method (*Gibbs*) with a naive Metropolis-Hasting algorithm (*MH*) and a Hamiltonian Monte Carlo (*HMC*) sampler (where ϵ and n_{step} are selected via a grid search, see appendix A.7) both implemented in Turing.jl (Ge et al., 2018), with a whitening transformation on the kernel matrix for better mixing. We draw 5 independent chains of 10000 samples for each algorithm. We compare crucial sampling diagnostics among different models: we give the autocorrelation between consecutive samples (lag 1) (as well as the autocorrelation plots for all lags in appendix A.7) to estimate the efficient sample size and the chain intercorrelation via the Gelman test (1 is the optimum) (Brooks and Gelman, 1998). The results are summarized in table 2.
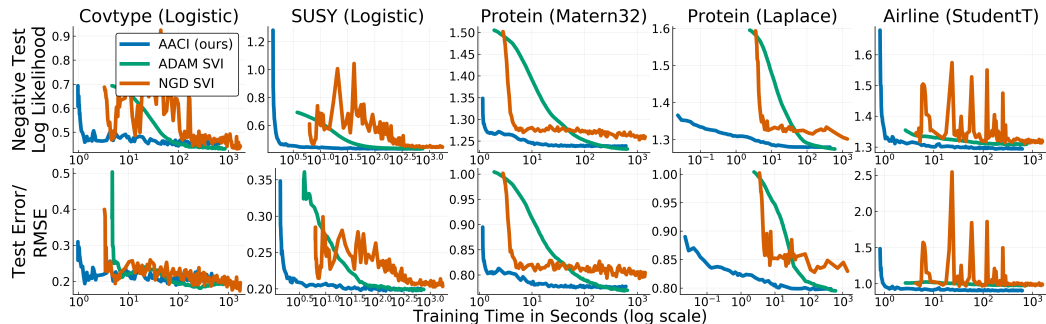
Figure 3. Test negative log-likelihood and test error (classification)/RMSE (regression) as a function of time for different likelihoods.

a) Matern 3/2 Likelihood on the Boston Housing dataset

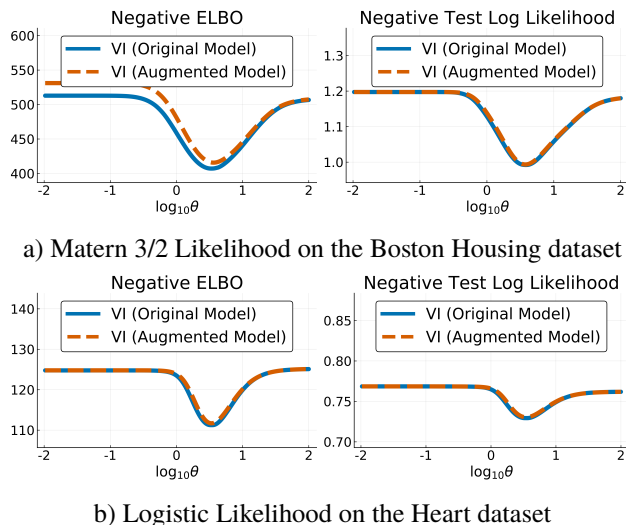b) Logistic Likelihood on the Heart dataset

Figure 2. Converged negative ELBO and averaged negative log-likelihood on a held-out dataset in function of the kernel lengthscale, training VI with and without augmentation.

We find that our method has a very low intrachain correlation leading to a high sample efficiency, as well as a low interchain correlation while still being faster than the HMC algorithm. It is even more evident for heavy-tailed likelihood like Student-T or Laplace where HMC can be of more trouble (Betancourt, 2017). Our approach is limited by the $\mathcal{O}(N^3)$ complexity for each sample.

5.2 Augmentation gap

To investigate the effect of augmenting the model when using variational inference, we train the original model using gradient descent and the augmented model until convergence. While we fix the kernel variance at 0.1, we vary the lengthscale θ from 10^{-2} to 10^2. We compare the converged ELBOs as well as the predictive performance on held-out test set. The results for the matern 3/2 and logistic are shown on figure 2, the other likelihoods are show in the appendix A.7. For both shown likelihoods, there is a visible ELBO gap between the augmented model and the original model. However the predictive performance is marginally

the same for both models.We can conclude that a potential difference in ELBO values does not affect the prediction performance.

5.3 Convergence speed

To scale our model to large datasets, we use the inducing points technique of Titsias (2009) and we use the stochastic gradient descent approach of Hoffman et al. (2013). We compare our variational approach (Algorithm 1) to using natural gradient descent, (Salimbeni et al., 2018) and ADAM (Hensman et al., 2015) both implemented in GPFlow (Matthews et al., 2017). For all methods we use 200 inducing points determined by k-means++ (Arthur and Vassilvitskii, 2007), minibatches of size 100 and we train the kernel hyperparameters using ADAM (Kingma and Ba, 2014), (the inducing points locations are fixed). We show the predictive performance in function of the training time for multiple likelihoods on figure 3.

Our method is up to two orders of magnitude faster than the state of the art. Moreover, we find that the optimization in our method is more stable (smooth decrease of the loss.

6 CONCLUSION

We proposed a new efficient inference method for GP models that have a super-Gaussian likelihood. Our method builds on an auxiliary variable augmentation that renders the model conditionally conjugate. We showed that in the augmented model, variational inference is up to two orders of magnitude faster and more stable than the state of the art. For small dataset, we proposed a Gibbs sampler that outperforms Hamiltonian Monte Carlo sampling. Previous methods that build on auxiliary variable augmentations (e.g. Wenzel et al., 2019) manually derived the augmentation and inference methods, whereas in our approach the whole procedure is fully automated and works for much more general class of models. Future work may aim on extending our approach to more general models by automatically constructing *hierarchical augmentations* inspired by Galy-Fajou et al. (2019) or Donner and Opper (2018).

References

Abate, J., Choudhury, G. L., and Whitt, W. (2000). An introduction to numerical transform inversion and its application to probability models. In *Computational probability*, pages 257–323. Springer.

Amari, S.-I. (1998). Natural gradient works efficiently in learning. *Neural computation*, 10(2):251–276.

Arthur, D. and Vassilvitskii, S. (2007). k-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 1027–1035. Society for Industrial and Applied Mathematics.

Beckers, T., Kulić, D., and Hirche, S. (2019). Stable gaussian process based tracking control of euler-lagrange systems. *Automatica*, (103):390–397.

Bernstein, S. et al. (1929). Sur les fonctions absolument monotones. *Acta Mathematica*, 52:1–66.

Betancourt, M. (2017). A conceptual introduction to hamiltonian monte carlo. *arXiv preprint arXiv:1701.02434*.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. springer.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877.

Brooks, S. P. and Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of computational and graphical statistics*, 7(4):434–455.

Cohen, A. M. (2007). *Numerical methods for Laplace transform inversion*, volume 5. Springer Science & Business Media.

Debnath, L. and Bhatta, D. (2014). *Integral transforms and their applications*. Chapman and Hall/CRC.

Devroye, L. (1986). *Nonuniform random variate generation*. Springer-Verlag.

Donner, C. and Opper, M. (2018). Efficient bayesian inference of sigmoidal gaussian cox processes. *The Journal of Machine Learning Research*, 19(1):2710–2743.

Dua, D. and Graff, C. (2017). UCI machine learning repository.

Eleftheriadis, S., Rudovic, O., Deisenroth, M. P., and Pantic, M. (2017). Gaussian process domain experts for modeling of facial affect. *IEEE Transactions on Image Processing*, 26(10):4697–4711.

Galy-Fajou, T., Wenzel, F., Donner, C., and Opper, M. (2019). Multi-class gaussian process classification made conjugate: Efficient inference via data augmentation. *Uncertainty in Artificial Intelligence (UAI)*.

Ge, H., Xu, K., and Ghahramani, Z. (2018). Turing: a language for flexible probabilistic inference. In *International Conference on Artificial Intelligence and Statistics, AISTATS*, pages 1682–1690.

Gneiting, T. (1997). Normal scale mixtures and dual probability densities. *Journal of Statistical Computation and Simulation*, 59(4):375–384.

Gneiting, T. (1999). Radial positive definite functions generated by euclid's hat. *Journal of Multivariate Analysis*, 69(1):88–119.

Hensman, J., Matthews, A., and Ghahramani, Z. (2015). Scalable variational gaussian process classification. *The Journal of Machine Learning Research*.

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

Jaakkola, T. S. and Jordan, M. I. (2000). Bayesian parameter estimation via variational methods. *Statistics and Computing*, 10(1):25–37.

Jylänki, P., Vanhatalo, J., and Vehtari, A. (2011). Robust gaussian process regression with a Student-t likelihood. *Journal of Machine Learning Research*, 12(Nov):3227–3257.

Khan, M. E. and Lin, W. (2017). Conjugate-computation variational inference: Converting variational inference in non-conjugate models to inferences in conjugate models. *International Conference on Artificial Intelligence and Statistics, AISTATS*.

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

Lawrence, N. D., Rattray, M., and Titsias, M. K. (2009). Efficient sampling for gaussian process inference using control variables. In *Advances in Neural Information Processing Systems*, pages 1681–1688.

Matthews, D. G., Alexander, G., Van Der Wilk, M., Nickson, T., Fujii, K., Boukouvalas, A., León-Villagrá, P., Ghahramani, Z., and Hensman, J. (2017). Gpflow: A gaussian process library using tensorflow. *The Journal of Machine Learning Research*, 18(1):1299–1304.

Merkle, M. (2014). Completely monotone functions: a digest. In *Analytic Number Theory, Approximation Theory, and Special Functions*, pages 347–364. Springer.

Mohamed, S., Rosca, M., Figurnov, M., and Mnih, A. (2019). Monte carlo gradient estimation in machine learning. *arXiv preprint arXiv:1906.10652*.

Palmer, J., Kreutz-Delgado, K., Rao, B. D., and Wipf,

D. P. (2006). Variational em algorithms for non-gaussian latent variable models. In *Advances in neural information processing systems*, pages 1059–1066.

Palmer, J. A. (2006). *Variational and scale mixture representations of non-Gaussian densities for estimation in the Bayesian linear model: Sparse coding, independent component analysis, and minimum entropy segmentation*. PhD thesis, UC San Diego.

Pandit, R. K. and Infield, D. (2018). Comparative analysis of binning and gaussian process based blade pitch angle curve of a wind turbine for the purpose of condition monitoring. *Journal of Physics: Conference Series*, 1102.

Polson, N. G., Scott, J. G., and Windle, J. (2013). Bayesian inference for logistic models using pólya–gamma latent variables. *Journal of the American statistical Association*, 108(504):1339–1349.

Polson, N. G., Scott, S. L., et al. (2011). Data augmentation for support vector machines. *Bayesian Analysis*, 6(1):1–23.

Rasmussen, C. E. (2003). *Gaussian processes in machine learning*. Springer.

Ridout, M. S. (2009). Generating random numbers from a distribution specified by its laplace transform. *Statistics and Computing*, 19(4):439.

Salimbeni, H., Eleftheriadis, S., and Hensman, J. (2018). Natural gradients in practice: Non-conjugate variational inference in gaussian process models. *roceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Schoenberg, I. J. (1938). Metric spaces and completely monotone functions. *Annals of Mathematics*, pages 811–841.

Titsias, M. (2009). Variational learning of inducing variables in sparse gaussian processes. In *Artificial Intelligence and Statistics*, pages 567–574.

Titsias, M. K., Lawrence, N., and Rattray, M. (2008). Markov chain monte carlo algorithms for gaussian processes. *Inference and Estimation in Probabilistic Time-Series Models*, 9.

Wang, C. and Blei, D. M. (2013). Variational inference in nonconjugate models. *Journal of Machine Learning Research*, 14(Apr):1005–1031.

Wenzel, F., Galy-Fajou, T., Deutsch, M., and Kloft, M. (2017). Bayesian nonlinear support vector machines for big data. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 307–322. Springer.

Wenzel, F., Galy-Fajou, T., Donner, C., Kloft, M., and Opper, M. (2019). Efficient gaussian process classification using Pòlya-gamma data augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 5417–5424.

Widder, D. V. (1946). *The Laplace transform*. Princeton university press.

Yeh, J. (2006). *Real analysis: theory of measure and integration second edition*. World Scientific Publishing Company.