Monotonic Gaussian Process Flows

Ivan Ustyuzhaninov* University of Tübingen

Ieva Kazlauskaite* University of Bath, Electronic Arts

Carl Henrik Ek University of Bristol

Neill D. F. Campbell University of Bath, Royal Society

Abstract

We propose a new framework for imposing monotonicity constraints in a Bayesian nonparametric setting based on numerical solutions of stochastic differential equations. We derive a nonparametric model of monotonic functions that allows for interpretable priors and principled quantification of hierarchical uncertainty. We demonstrate the efficacy of the proposed model by providing competitive results to other probabilistic monotonic models on a number of benchmark functions. In addition, we consider the utility of a monotonic random process as a part of a hierarchical probabilistic model; we examine the task of temporal alignment of time-series data where it is beneficial to use a monotonic random process in order to preserve the uncertainty in the temporal warpings.

1 INTRODUCTION

Monotonic regression is a task of inferring the relationship between a dependent variable y and an independent variable x when it is known that the relationship $y = f(x)$ is monotonic. Monotonic functions (and monotonic random processes) have previously been studied in areas as diverse as physical sciences for estimating the temperature of a cannon barrel over time [Lavine and Mockus, 1995], marine biology for surveying of fauna on the seabed of the Great Barrier Reef [Hall and Huang, 2001], geology for chronology of sediment samples [Haslett and Parnell, 2008], public health for relating obesity and body fat [Dette and Scheder, 2006], sociology for relating education, work experience and salary [Dette and Scheder, 2006], design of computer networking systems [Golchi et al., 2015], economics for estimating personal income [Canini et al., 2016], insurance for predicting mortality rates [Durot and Lopuhaä, 2018], biology of establishing the diagnostic value of bio-markers for Alzheimer's disease and for trajectory estimation in brain imaging [Lorenzi et al., 2019, Nader et al., 2019], meteorology for estimation of wind-induced under-catch of winter precipitation [Kim et al., 2018] and others.

Monotonicity also appears in the more general context of hierarchical models where we want to transform a (simple and typically stationary) input distribution to a (complicated and non-stationary) data distribution. More specifically, monotonicity constraints have been used in hierarchical models with warped inputs, for example, in Bayesian optimisation of non-stationary functions [Snoek et al., 2014] and in mixed effects models for temporal warps of time-series data [Kaiser et al., 2018, Kazlauskaite et al., 2019, Raket et al., 2016].

Extensive study by the statistics [Ramsay, 1988, Sill and Abu-Mostafa, 1997] and machine learning communities [Riihimäki and Vehtari, 2010, Andersen et al., 2018] has resulted in a variety of frameworks. While many traditional approaches use constrained parametric splines, they are not sufficiently expressive and, typically, do not include prior beliefs about the characteristics of the underlying function (such as smoothness). Consequently, many contemporary methods consider monotonicity in the context of continuous random processes, mostly based on Gaussian processes (GPs) [Rasmussen and Williams, 2005]. As a nonparametric Bayesian model, a GP is an attractive foundation on which to build flexible and theoretically sound models with well-calibrated estimates of uncertainty and automatic complexity control. However, imposing monotonicity constraints on a GP has proven to be problematic [Lin and Dunson, 2014, Riihimäki and Vehtari, 2010] as it requires both formulating a prior that is monotonic as well as constraining the (predictive) posterior to be monotonic.

* Equal contribution

Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108.

This is particularly challenging as monotonicity is a global property, implying that the function values are correlated for all inputs, irrespective of the lengthscale of the covariance [Andersen et al., 2018].

In this work we propose a novel nonparametric Bayesian model of monotonic functions that is based on the recent work on differential equations (DEs). At the heart of such models is the idea of approximating the derivatives of a function rather than studying the function directly. DE models have gained a lot of popularity recently and they have been successfully applied in conjunction with both neural networks [Chen et al., 2018] and GPs [Heinonen et al., 2018, Yildiz et al., 2018a, Yildiz et al., 2018b]. We consider a recently proposed framework, called differential GP flows [Hegde et al., 2019], that performs classification and regression by learning a stochastic differential equation (SDE) transformation of the input space. It admits an expressive yet computationally convenient parametrisation using GPs.

Utilising the uniqueness theorem for the solutions of SDEs [Øksendal, 1992], we formulate a novel stochastic random process that is guaranteed to be monotonic. We show that, unlike some of the previous work on monotonic random processes, the proposed approach is guaranteed to lead to monotonic samples from the model (defined as a flow field), and it performs competitively on a set of regression benchmarks.

Furthermore, we study an illustrative example of a hierarchical, two-layer model where the first layer corresponds to a smooth monotonic warping of time and the second layer corresponds to a sequence of time-series observations. The overall goal in such a problem is to learn the warpings of the inputs such that the unwarped versions of the sequences are temporally aligned. While such models typically rely on parametric transformations for the temporal warpings [Kazlauskaite et al., 2019], we show how the estimation of uncertainty leads to a model that is more informative and more interpretable than the previous approach. To achieve this, we further make use of the recent advances in variational inference for deep GPs [Ustyuzhaninov et al., 2019] to capture the compositional uncertainty present in the hierarchical model.

2 RELATED WORK

Splines Many classical approaches to monotonic regression rely on spline smoothing: given a basis of monotone spline functions, the underlying function is approximated using a non-negative linear combination of these basis functions and the monotonicity constraints are satisfied in the entire domain [Wahba, 1978] by construction. For example, Ramsey [Ramsay, 1998] considers a family of functions defined by the differential equation $D^2 f = \omega D f$ which contains the strictly monotone twice differentiable functions, and approximates ω using a basis of M-splines and I-splines. Shively *et al.* [Shively et al., 2009] consider a finite approximation using quadratic splines and a set of constraints on the coefficients that ensure isotonicity at the interpolation knots. The use of piecewise linear splines was explored by Haslett and Parnell [Haslett and Parnell, 2008] who use additive i.i.d. gamma increments and a Poisson process to locate the interpolation knots; this leads to a process with a random number of piecewise linear segments of random length, both of which are marginalised analytically. Further examples of spline based approaches rely on cubic splines [Wolberg and Alfy, 2002], mixtures of cumulative distribution functions [Bornkamp and Ickstadt, 2009] and an approximation of the unknown regression function using Bernstein polynomials [Curtis and Ghosh, 2011].

Gaussian process A GP is a stochastic process which is fully specified by its mean function, $\mu(\mathbf{x})$, and its covariance function, $k(\mathbf{x}, \mathbf{x}')$, such that any finite set of random variables have a joint Gaussian distribution [Rasmussen and Williams, 2005]. GPs provide a robust method for modeling non-linear functions in a Bayesian nonparametric framework; ordinarily one considers a GP prior over the function and combines it with a suitable likelihood to derive a posterior estimate for the function given data. The nonparametric nature of the GP means that, unlike the parametric counterparts, it adapts to the complexity of the data.

Monotonic Gaussian processes A common approach is to ensure that the monotonicity constraints are satisfied at a finite number of input points. For example, Da Veiga and Marrel [Da Veiga and Marrel, 2012] use a truncated multi-normal distribution and an approximation of conditional expectations at discrete locations, while Maatouk [Maatouk, 2017] and Lopez-Lopera *et al.* [Lopez-Lopera et al., 2019] proposed finite-dimensional approximations based on deterministic basis functions evaluated at a set of knots. Another popular approach proposed by Riihimäki and Vehtari [Riihimäki and Vehtari, 2010] is based on including the derivatives information at a number of input locations by forcing the derivative process to be positive at these locations. Extensions to this approach include both adapting to new application domains [Golchi et al., 2015, Lorenzi et al., 2019, Siivola et al., 2016] and proposing new inference schemes [Golchi et al., 2015]. However, these approaches do not guarantee monotonicity as they impose constraints at a finite number of points only. Lin and Dunson [Lin and Dunson, 2014] propose another GP based approach that relies on projecting sample paths from a GP to the space of monotone functions using pooled adjacent violators algorithm which does not impose smoothness. Furthermore, the projection operation complicates the inference of the parameters of the GP and produces distorted credible intervals. Lenk and Choi [Lenk and Choi, 2017] design shape restricted functions by enforcing that the derivatives of the functions are squared Gaussian processes and approximating the GP using a series expansion with the Karhunen-Loève representation and numerical integration. Andersen *et al.* [Andersen et al., 2018] follow a similar approach, in which the derivatives of the functions are assumed to be compositions of a GP and a non-negative function; in the following we refer to this method as transformed GP.

3 BACKGROUND

We now discuss the SDE framework we build upon for our monotonic random process. Any random process can be defined through its finite-dimensional distribution [Øksendal, 1992]. This implies that modelling the observations $\{f(x_n)\}_{n=1}^N$ with trajectories of such a process requires their definition through the finite-dimensional joint distributions $p(f(x_1), \dots, f(x_N))$. Constraining the functions to be monotonic necessitates choosing a family of joint probability distributions that satisfies the monotonicity constraint:

$$\begin{aligned} p(f(x_1), \dots, f(x_N)) &= 0, \\ &\text{unless } f(x_1) \leq \dots \leq f(x_N). \end{aligned} \tag{MC}$$

This could be achieved by truncating a standard joint distribution (*e.g.* Gaussian) but inference in such models is computationally challenging [Maatouk, 2017]. Another approach is to define a random process to have monotone trajectories by construction (*e.g.* Compound Poisson process) but this often requires making simplifying assumptions on the trajectories (and therefore on $\{f(x)\}$). In contrast, we use solutions of SDEs to define a random process with monotonic trajectories by construction while avoiding strong simplifying assumptions.

3.1 Gaussian process flows

SDE solutions Our model builds on the general framework for modelling functions as SDE solutions introduced in [Hegde et al., 2019]. Consider the following SDE:

$$\begin{aligned} dS(t, \omega; x) &= \mu\left(S(t, \omega; x), t\right) \, dt \\ &+ \sqrt{\sigma\left(S(t, \omega; x), t\right)} \, dW(t, \omega) \end{aligned} \tag{1}$$

where $W(t, \omega)$ is the Wiener process. The solution of this SDE is a stochastic process $S(t, \omega; x)$ which is a function of three arguments: the time t, the initial value x at time $t = 0$, and the element $\omega \in \Omega$ of the underlying sample space Ω.[1]

For a fixed time $t = T$, the corresponding SDE solution $S(T, \omega; x)$ is a random variable that depends on the initial condition x. Therefore, there exists a mapping of an arbitrary initial condition to this solution at time T: $x \mapsto S(T, \omega; x)$ and the distribution of the SDE solutions induces a distribution over such mappings (similar to GPs, for example). The family of such distributions is parametrised by functions $\mu\left(S(t, \omega; x), t\right)$ (drift) and $\sigma\left(S(t, \omega; x), t\right)$ (diffusion), which are defined in [Hegde et al., 2019] using a sparse Gaussian process [Titsias, 2009].

Flow GP Consider a zero-mean, single-output GP $g \sim \mathcal{GP}(0,\, k(\cdot, \cdot))$, which is a function of two arguments: a space variable s and time t. We specify the GP via a set of M inducing outputs $\mathbf{U} = \{U_m\}_{m=1}^M$, $U_m \in \mathbb{R}$, corresponding to inducing input locations $\mathbf{Z} = \{\mathbf{z}_m\}_{m=1}^M$, $\mathbf{z}_m \in (\{s\} \times \{t\}) = \mathbb{R}^2$, similarly to [Titsias, 2009]. The predictive posterior distribution of such a GP evaluated at a spatio-temporal point (s, t) is as follows:

$$\begin{aligned} &p(g(s,t) \mid \mathbf{U}, \mathbf{Z}) \sim \mathcal{N}(\widetilde{\mu}(s,t), \widetilde{\Sigma}(s,t)) \\ &\quad \widetilde{\mu}(s,t) = \mathbf{K}_{(s,t),\mathbf{z}} \, \mathbf{K}_{\mathbf{zz}}^{-1} \, \mathbf{U}, \\ &\quad \widetilde{\Sigma}(\mathbf{x},t) = k((s,t),(s,t)) - \mathbf{K}_{(s,t),\mathbf{z}} \, \mathbf{K}_{\mathbf{zz}}^{-1} \, \mathbf{K}_{\mathbf{z},(s,t)}, \end{aligned} \tag{2}$$

where the covariance matrix $\mathbf{K}_{\mathbf{ab}} := k(\mathbf{a}, \mathbf{b})$. We define the SDE drift and diffusion functions to be $\mu(S(t, \omega; x), t) := \widetilde{\mu}(S(t, \omega; x), t)$ and $\sigma(S(t, \omega; x), t) := \widetilde{\Sigma}(S(t, \omega; x), t)$ implying that (1) is completely defined by the GP g and its set of inducing points $\{\mathbf{U}, \mathbf{Z}\}$. Similarly to [Hegde et al., 2019], the joint density of a single path then is (neglecting $\mathbf{Z}$ for clarity):

$$\begin{aligned} p(y, S(T, \omega; x), g, \mathbf{U}) = &\underbrace{p(y \mid S(T, \omega; x))}_{\text{likelihood}} \\ &\underbrace{p\left(S(T, \omega; x) \mid g\right)}_{\text{SDE}} \; \underbrace{p\left(g \mid \mathbf{U}\right) p\left(\mathbf{U}\right)}_{\text{GP prior of } g(s,t)}. \end{aligned} \tag{3}$$

Inference Inferring $\mathbf{U}$ is intractable in closed form, hence the posterior of $\mathbf{U}$ is approximated by a variational distribution $q(\mathbf{U}) \sim \mathcal{N}(\mathbf{m}, \mathbf{S})$, the parameters of which (and the inducing inputs $\mathbf{Z}$) are optimised by maximising the marginal likelihood lower bound $\mathcal{L}$:

$$\begin{aligned} \log p(\mathcal{D}) \geq \mathcal{L} := &- \mathrm{KL}[\, q(\mathbf{U}) \,||\, p(\mathbf{U}) \,] \\ &+ \mathbb{E}_{q(\mathbf{U})} \mathbb{E}_{p(S(T,\omega;x) \mid \mathbf{U})} [\, \log p\left(y \mid S(T, \omega; x)\right)]. \end{aligned} \tag{4}$$

[1] Typically, the dependencies on x and ω are omitted, denoting the stochastic process as S_t, however, these dependencies are crucial for our construction of the monotonic flow model, thus we explicitly keep them in the notation.

The expectation $\mathbb{E}_{p(S(T,\omega;\mathbf{x})\,|\,\mathbf{U})}$ is approximated by sampling the numerical approximations of the SDE solutions. This is particularly convenient to do with $\mu\left(S(t,\omega;x),t\right)$ and $\sigma\left(S(t,\omega;x),t\right)$ defined as parameters of a GP posterior as sampling such an SDE solution only requires generating samples from the posterior of the GP given the inducing points $\mathbf{U}$ (see [Hegde et al., 2019] for details). The first term in (4) is a KL divergence between two Gaussian distributions available in closed form.

4 MONOTONIC GAUSSIAN PROCESS FLOW

We now describe our proposed random process with monotonic trajectories. Assuming N one-dimensional initial conditions (denoted jointly as $\mathbf{x} = (x_1, \ldots, x_N) \in \mathbb{R}^N$), we use the SDE solution mapping $\mathbf{x} \mapsto S(T,\omega;\mathbf{x}) := (S(T,\omega;x_1), \ldots, S(T,\omega;x_N))$ as our model of monotonic function. We begin with an intuitive discussion of why $S(T,\omega;\mathbf{x})$ is a monotonic function of $\mathbf{x}$ using a fluid flow field analogy.

1. A general ordinary smooth DE $du(t) = \phi(u)dt$ may be thought of as a fluid flow field. Its solutions $u(t,x_1), \ldots, u(t,x_n)$ corresponding to the initial values $x_1, \ldots, x_n$ are trajectories or streams of particles in this field starting at these initial values. A fundamental property of such flows is that one can never cross the streams of the flow field.[2] Therefore, if particles are evolved simultaneously under a flow field their ordering cannot be permuted; this gives rise to a monotonicity constraint.
2. A stochastic differential equation, however, introduces random perturbations into the flow field so particles evolving *independently* could jump across flow lines and change their ordering. However, a single, coherent draw from the SDE (corresponding to an individual realisation of the paths $W(\cdot,\omega)$) will always produce a valid flow field (the flow field will simply change between draws). Thus, particles evolving jointly under a single draw will still evolve under a valid flow field and therefore never permute.

4.1 SDE solutions are monotonic functions of initial values

The joint distribution $p\left(S(T,\omega;x_1), \ldots, S(T,\omega;x_N)\right)$ of solutions of the SDE in (1) with initial values $x_1 \leq \ldots \leq x_N$ satisfies (MC).

This follows from a general result that SDE solutions $S(t,\omega;x)$ are unique and continuous under certain regularity assumptions for any initial value x (see, for example, Theorem 5.2.1 in [Øksendal, 1992]). Specifically, a random variable $S(t,\omega;x)$ is a unique and continuous function of t for any element of the sample space $\omega \in \Omega$. Using this result we conclude that if we have two initial conditions x and x' such that $x \leq x'$, the corresponding solutions at some time T also obey this ordering, *i.e.* $S(T,\omega;x) \leq S(T,\omega;x')$ for $\omega \in \Omega$. Indeed, were that not the case, the continuity of $S(t,\omega;x)$ as a function of t implies that there exists some $0 \leq t_c \leq T$ such that $S(t_c,\omega;x) = S(t_c,\omega;x')$ (*i.e.* the trajectories corresponding to initial values x and x' cross), resulting in two different solutions of the SDE for the initial condition $x_c := S(t_c,\omega;x) = S(t_c,\omega;x')$ (namely $S(T-t_c,\omega;x_c) = S(T,\omega;x)$ and $S(T-t_c,\omega;x_c) = S(T,\omega;x')$), violating the uniqueness result.

The above argument assumes a fixed flow field (defined by the drift and the diffusion functions) and a fixed Wiener realisation (corresponding to $W(\cdot,\omega)$); it implies that individual solutions (*i.e.* solutions to a single draw) of the SDEs at a fixed time T, $S(T,\omega;x)$, are monotonic functions of the initial conditions, and hence define a random process with monotonic trajectories. The actual prior distribution of such trajectories depends on the exact form of the functions $\mu\left(S(t,\omega;x),t\right)$ and $\sigma\left(S(t,\omega;x),t\right)$ in (1) (*e.g.* if $\sigma(S(t,\omega;x),t) = 0$, the SDE is an ordinary DE and $S(T,\omega;x)$ is a deterministic function of $\mathbf{x}$ independent of ω, meaning that the prior distribution consists of a single monotonic function). Prior distributions over $\mu\left(S(t,\omega;\mathbf{x}),t\right)$ and $\sigma\left(S(t,\omega;\mathbf{x}),t\right)$ thus induces priors over the monotonic functions $S(T,\omega;\mathbf{x})$, and inference in this model consists of computing the posterior distribution of these functions conditioned on the observed noisy sample from a monotonic function. For details of the numerical solution of the SDE, see supplementary material A.1.

4.2 Notable differences to Hedge et al.

1. In [Hegde et al., 2019], a regular GP is placed on top of the SDE solutions $S(T,\omega;\mathbf{x})$, so that $p\left(\mathbf{y}\,|\,S(T,\omega;\mathbf{x})\right)$ is a GP with a Gaussian likelihood in (4). In contrast, since we are modelling monotonic functions and $S(T,\omega;\mathbf{x})$ are monotonic functions of $\mathbf{x}$, we define $p\left(Y\,|\,S(T,\omega;\mathbf{x})\right)$ to be directly the likelihood
$$p\left(\mathbf{y}\,|\,S(T,\omega;\mathbf{x})\right) = \mathcal{N}(\mathbf{y}\,|\,S(T,\omega;\mathbf{x}), \sigma^2 I), \quad (5)$$
where $\mathbf{y}$ is a vector of observations sampled from an underlying unknown monotonic function $f(\mathbf{x})$.
2. The argument in this section assumes a fixed flow field (defined by the drift and the diffusion functions) and a fixed Wiener realisation (denoted by ω). Thus, a critical difference in our inference

[2] Also an important safety tip to avoid total protonic reversal [Spengler, 1984].

procedure is that at every iteration of the numerical SDE solver, we jointly sample the increments $\Delta\mathbf{x}$ in the flow field using (2). This ensures that they are taken from the same instantaneous realisation of the stochastic flow field and hence the monotonicity constraint is satisfied.

5 EXPERIMENTS

First, we test the monotonic flow model on the task of estimating monotonic curves from noisy observations (in high and low data regimes) before investigating the quantification of uncertainty.

Regression We use a set of 6 benchmark functions from previous studies [Lin and Dunson, 2014, Maatouk, 2017, Shively et al., 2009]. Three examples of the functions are shown in Fig. 1; the exact equations are in the Supplement A.5. The training data is generated by evaluating these functions at N equally spaced points and adding i.i.d. Gaussian noise $\varepsilon_n \sim \mathcal{N}(0, 1)$. We note that many real-life datasets that benefit from monotonicity constraints have similar trends and high levels of noise (*e.g.* [Haslett and Parnell, 2008, Curtis and Ghosh, 2011, Kim et al., 2018]). Following the literature, we used the root-mean-square-error (RMSE) to evaluate performance.

100 data points Table 1 in the Supplement A.8 provides the results obtained by fitting different monotonic models to data sets containing $N = 100$ points. As baselines we include: GPs with monotonicity information [Riihimäki and Vehtari, 2010][3], transformed GPs [Andersen et al., 2018][4], and other results reported in the literature. We report the RMSE means and the SD from 20 trial runs with different random noise samples and show example fits in the bottom row of Fig. 1. This figure contains the means of the predicted curves from 10 trials with the best parameter values (each trial contains a different sample of standard Gaussian random noise). We plot samples as opposed to the mean and the SD as, due to the monotonicity constraint, samples are more informative than sample statistics. The parameter values we cross-validated over are detailed in the Supplement A.6.

Overall, our method performs very competitively, achieving the best results on 3 functions and being within a standard deviation of the best result on all others. We note that the training data contains a lot of observational noise (see Fig. 1), thus using prior monotonicity assumptions significantly improves results over a regular GP.

[3]Implementation available from `https://research.cs.aalto.fi/pml/software/gpstuff/`.

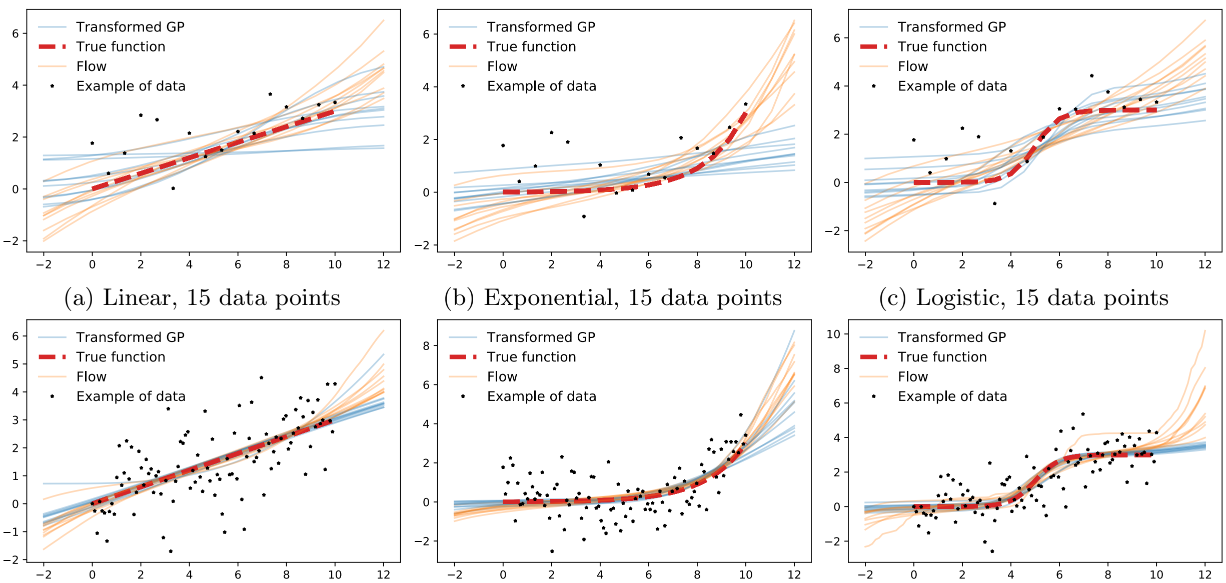
[4]Implementation provided in personal communications.

15 data points In Table 2 in the Supplement A.8 and Fig. 1 (top row) we provide a comparison of the flow and the transformed GP in a setting when only $N = 15$ data points are available. Our fully nonparametric model is able to recover the structure in the data significantly better than the Transformed GP which usually reverts to a nearly linear fit on all functions. This might be explained by the fact that the Transformed GP is a parametric approximation of a monotonic GP, and the more parameters included, the larger the variety of the functions it can model. However, estimating a large (w.r.t. dataset size) number of parameters is challenging given a small set of noisy observations. The monotonic flow tends to underestimate the value of the function on the left side of the domain and overestimate the value on the right. The mean of our prior of the monotonic flow with a stationary flow GP kernel is an identity function, so given a small set of noisy observations, the predictive posterior mean quickly reverts to the prior distribution near the edges of the data interval.

Uncertainty quantification in monotonic random processes In standard (non-monotonic) regression, GPs are used as the gold standard for the quantification of uncertainty [Foong et al., 2019]. However, directly comparing the confidence intervals of a monotonic random process to a standard GP is misleading due to the additional constraints of monotonicity which lead to tighter confidence intervals as fewer explanations (functions) are compatible with the observed data. Fig. A2 illustrates the shrinking of the confidence intervals for monotonic random processes in comparison to a standard (unconstrained) GP. As a baseline, we fit a standard GP (Fig. A2a) and consider only those samples from the posterior which are monotonic increasing in the domain in which we perform extrapolation ($[-5, 5]$); these samples along with their mean and 2 SD away from the mean are shown in Fig. A2b[5]. The GP with monotonicity information (Fig. A2c) is not able to guarantee that the samples are monotonic, especially in parts of the domain away from the data, while the transformed GP (Fig. A2d) tends to underestimate the uncertainty, potentially due to the Dirichlet conditions imposed on the boundaries of the domain. Meanwhile, the uncertainty estimates of our proposed monotonic flow are comparable to the baseline (i.e. the monotonic samples from a standard GP) during extrapolation and samples from the flow are guaranteed monotone.

An alternative visualisation of the flow model involves

[5]We note that plotting the error bars using a Gaussian density may be misleading in monotonic regression as the samples from such process may not be symmetric around the mean, especially when the data are nearly constant, which can be seen by looking at the samples.



(a) Linear, 15 data points (b) Exponential, 15 data points (c) Logistic, 15 data points

(d) Linear, 100 data points (e) Exponential, 100 data points (f) Logistic, 100 data points

Figure 1: Mean fits for 10 trials with different random noise as estimated by the flow and the transformed GP [Andersen et al., 2018] (the noise samples are identical for both methods; we plot the data from one trial).

looking at the streamlines of the input values as a function of time (see Fig. 2). The streamlines may be visualised as one coherent draw from the flow (shown at the top of Fig. 2), or as independent samples at a given value of the inputs (shown in the middle figures of Fig. 2). The latter also help visualise the uncertainty in the model as these samples show the range of possible outputs $S(T, \omega; \mathbf{x})$ for a given input location $\mathbf{x}$.

The mean and variance of the inducing points in the flow GP depend on their number M as follows: given few inducing points, they are typically optimised to be located close to the observations so that the resulting model fits the observations well (with low estimated observational noise). Meanwhile, given a large number of inducing points, some are used to fit the data well while others are placed in regions with no observations (see, for example, the regions in between the data ($[-2, 2]$) in Fig. 2b) and optimised to have higher variance $\mathbf{S}$ in those regions. We note that the uncertainty estimates in the monotonic flow model do not depend much on the number of inducing points: the estimates are nearly identical for $M > 5$ while for this data $M = 5$ may not be enough to explain the data well, hence the observational noise gets overestimated, also resulting in higher variance in extrapolation. Fig. 3a shows how the uncertainty estimates for this data set depend on the number of inducing points. Similarly, Fig. 3b details the dependence on the flow time T [Hegde et al., 2019]; longer flow time ($T \geq 10$) results in more extreme warpings and thus higher uncertainty at the observations with overestimated observational noise.

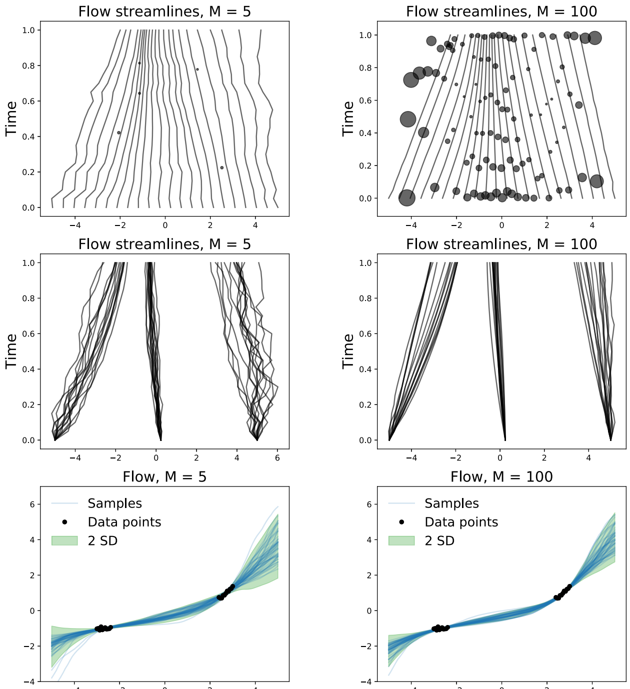
6 ALIGNMENT APPLICATION

A monotonic constraint in the first layer is desirable in mixed effects models where the first layer corresponds to a warping of space or time that does not allow permutations. We consider an application of the monotonic random process as an integral part of a model designed to align multiple temporal sequences of observations. This problem is introduced in detail in [Kazlauskaite et al., 2019] and here we provide a short summary and the baseline model. We change notation to match [Kazlauskaite et al., 2019]; for the alignment application, $g(\cdot)$ refers to the monotonic function, specified using the monotonic Gaussian process flow model, whereas $f(\cdot)$ is now an arbitrary function.

Assume we are given some time-series data with inputs $\mathbf{x} \in \mathbb{R}^N$ and J output sequences $\{\mathbf{y}_j \in \mathbb{R}^N\}_{j=1}^J$. We know that there are multiple underlying functions that generated this data, say K such functions, $f_k(\cdot)$, and the observed data were generated by warping the temporal inputs to the true functions using a monotonic warping function $g_j(\mathbf{x})$, such that:

$$\mathbf{y}_j = f_k\big(g_j(\mathbf{x})\big) + \epsilon_j. \tag{6}$$

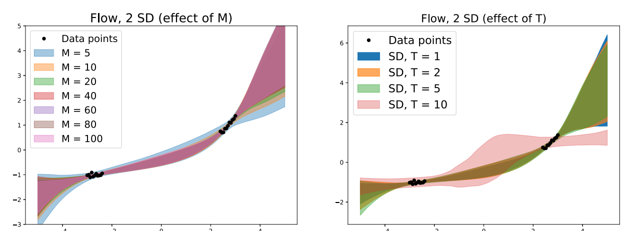
where $\epsilon_j \sim \mathcal{N}(0, \beta^{-1} I_N)$ is observation noise. Then the corresponding, latent, sequences that are not corrupted by the temporal warp (*i.e.* the aligned versions of $\mathbf{y}_j$) are $\mathbf{f}_j := f_k(\mathbf{x})$. The functions $f_j(\cdot)$ are modelled jointly using a GP and the joint conditional likelihood for each



(a) Flow streamlines for 5 inducing points.

(b) Flow streamlines for 100 inducing points.

Figure 2: A coherent sample (top) and a set of independent samples at three chosen input locations (middle) from a fitted flow (bottom). The circles (top figures) show the location $\mathbf{m}$ of the inducing points and are scaled by their (relative) variance $\mathbf{S}$.



(a) Flow comparison for $M = 5, 50, 100$.

(b) Flow comparison for $T = 1, 2, 5, 10$.

Figure 3: Effect of the number M of inducing points and the total flow time T on the estimated uncertainty (coloured regions correspond to 2 SD away from the mean of the samples from the flow). Results for 5 random trials.

pair of sequences, $\mathbf{y}_j$ and $\mathbf{f}_j$, is:

$$p\left(\begin{bmatrix}\mathbf{f}_j\\ \mathbf{y}_j\end{bmatrix}\middle|\, \mathbf{g}_j, \mathbf{x}, \theta_j\right) \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} k_{\theta_j}(\mathbf{x},\mathbf{x}) & k_{\theta_j}(\mathbf{x},\mathbf{g}_j)\\ k_{\theta_j}(\mathbf{g}_j,\mathbf{x}) & k_{\theta_j}(\mathbf{g}_j,\mathbf{g}_j)+\beta_j^{-1}\end{bmatrix}\right) \tag{7}$$

where $\mathbf{g}_j := g_j(\mathbf{x})$ are finite-dimensional realisations of the warping function, and θ_j includes the parameters of the GP that models function $f_j(\cdot)$ and the parameters of the warping function $g_j(\cdot)$. The task is then to

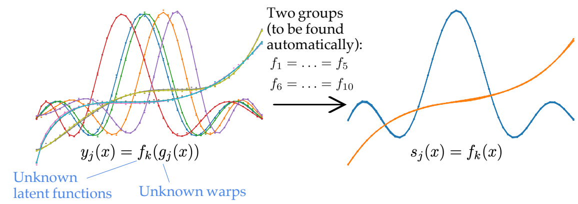


Figure 4: Illustration of the alignment problem.

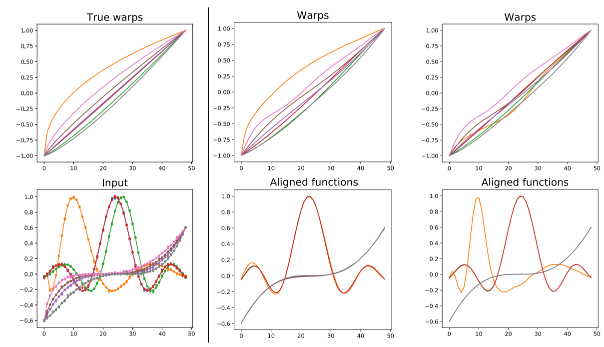
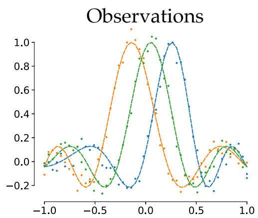


Figure 5: The observations (bottom left) were generated by applying warping functions (top left) to a sinc or cubic function ($K = 2$). The point estimates in the alignment model of [Kazlauskaite et al., 2019] will return one of the two possible solutions (middle and right columns). One solution (middle) aligns all the sequences into two groups but uses more extreme warps (note the orange warp) while the other solution (right) assigns the orange sequence to a new cluster (and thus uses an identity warp for this sequence). Both are plausible given our priors on the warps $g_j(\cdot)$ and the functions $f_j(\cdot)$, hence a preferred model would preserve the uncertainty about the warps and the cluster assignments and capture the full range of possible solutions.

learn the latent functions $f_k(\cdot)$ and the warps $g_j(\cdot)$ such that the versions of these function which are not corrupted by the warp, $\mathbf{f}_j$, are aligned as well as possible. Note that the number K of distinct functions $f_k(\cdot)$ is unknown must also be inferred from the data. This is achieved by formulating an alignment objective that pushes the uncorrupted sequences $\mathbf{f}_j$ into K groups where each group corresponds to one latent function $f_k(\cdot)$, and the sequences within each group are aligned to each other. If the warps fully explain the differences between the sequences within each group, then each group contains a single sequence $\mathbf{f}_k(\cdot)$ (or, equivalently, all the sequences within a group coincide); see Fig. 6.

Previously, [Kazlauskaite et al., 2019] proposed a probabilistic alignment objective based on a GP latent variable model (GP-LVM) [Lawrence, 2004] that aligns sequences within groups. A GP-LVM is a generative model that is often used as a dimensionality reduction technique to uncovers the latent structure in the data by constructing a low dimensional manifold, and using independent GPs as mappings from a latent space to an observed space. In a GP-LVM, GPs are taken to be



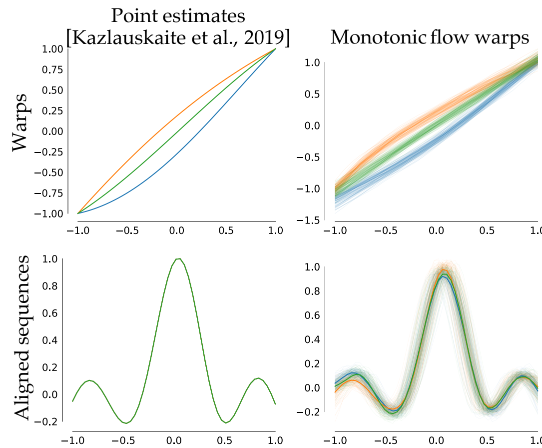


Figure 6: Given noisy observations of warped sequences, we compare the uncertainty in the warps for the proposed model (with and without correlations) and for [Kazlauskaite et al., 2019].

independent across the features (columns) of the data $\mathbf{F} = \{\mathbf{f}_j\}_{j=1}^J \in \mathbb{R}^{J \times N}$, and the likelihood function is:

$$p(\mathbf{F} \mid \mathbf{v}) = \prod_{n=1}^N \mathcal{N}(\mathbf{F}_{:,n} \mid 0, k(\mathbf{v}, \mathbf{v}) + \hat{\beta}^{-1} I_J) \qquad (8)$$

where $\mathbf{v} = \{v_j\}_{j=1}^J$ are the latent variables for each sequence, the GP covariance is $k(\mathbf{v}, \mathbf{v})$ and $\hat{\beta}$ is a noise precision. Typically, a GP-LVM contains a prior on the latent variables $p(\mathbf{v}) \sim \mathcal{N}(\mathbf{0}, \sigma_v^2 I_J)$. This encourages the latent variables to be placed close to the origin in the latent space if the observed data can be explained by a single cluster in a latent space; otherwise a minimal number of clusters of $\mathbf{v}$ in the latent space will be made under the Bayesian prior encouraging sparsity. Furthermore, the stationary kernel in the GPs that map from the latent space to the observed space depend only on the distance between the latent location v_j, which means that the further each latent location is from the others, the less correlated the corresponding GP outputs $\mathbf{f}_j$. Specifically, this behaviour is controlled by the kernel length-scale which allows to reduce correlation between groups of sequences while maintaining strong correlation among sequences within each group.

The two objectives of (7) and (8) are combined to learn the GPs for $f_j(\cdot)$, the warps $g_j(\cdot)$ and the number of underlying clusters K. The baseline [Kazlauskaite et al., 2019] uses MAP estimates for the fitting of the GPs in (7) and for the GP-LVM in (8). This leads to point estimates of the warps $\mathbf{g}_j$ and, consequently, does not retain any information about the uncertainty of cluster assignments. Fig. 5 illustrates the limitation of using point estimates; the observed data can be explained in multiple different ways which cannot be uncovered using point estimates.

Uncertainty in monotonic warps We demonstrate the ability of our monotonic random process to capture the uncertainties in the warps and the cluster assignments in the alignment model. In order to preserve the compositional uncertainty in both the warping functions $g_j(\cdot)$ and the latent functions $f_j(\cdot)$, we introduce correlations between the variational distributions in the two layers, $g(\cdot)$ and $f(\cdot)$ using the inference scheme detailed in [Ustyuzhaninov et al., 2019]. Fig. 6 illustrates this phenomenon, and compares the uncertainty in the warps for the original point estimate [Kazlauskaite et al., 2019], the monotonic flow and the flow with correlations between the samples from the warp and the function f. The flow captures a range of different possible warpings that are consistent with our prior (which favours solutions that are close to an identity warp) and also fits and aligns the data well. An additional example with bi-modal behaviour, as in Fig. 5, is given in Fig. A1 in the Supplement.

7 CONCLUSION

We have proposed a novel nonparametric model of monotonic functions based on a random process with monotonic trajectories that confers improved performance over the state-of-the-art as well as preferable theoretical properties. Many real-life regression tasks deal with functions that are known to be monotonic, and explicitly imposing this constraint helps uncover the structure in the data, especially when the observations are noisy or data are scarce. We have also demonstrated that the proposed construction can be used as part of a complex alignment model where the uncertainty estimates provide a more informative model and help uncover structures in the data that are not captured by the existing models. More broadly, with additional mid-hierarchy marginal information or domain specific knowledge of compositional priors, *e.g.* [Kaiser et al., 2018], hierarchical models may necessitate a composition of (injective) monotonic mappings for all but the output layer. This advocates the study of monotonic functions, which can represent a wide variety of transformations and hence serve as a general purpose first layer in a hierarchical model, especially, when the function is known to be non-stationary.

Acknowledgments

This work has been supported by EPSRC CDE (EP/L016540/1) and CAMERA (EP/M023281/1) grants as well as the Royal Society. The authors are grateful to Markus Kaiser and Garoe Dorta for their insight and feedback on this work, and to Michael Andersen for sharing the implementation of Transformed GPs. IK would like to thank the Frostbite Physics team at EA.

References

[Andersen et al., 2018] Andersen, M. R., Siivola, E., Riutort-Mayol, G., and Vehtari, A. (2018). A non-parametric probabilistic model for monotonic functions. *"All Of Bayesian Nonparametrics" Workshop at NeurIPS*.

[Bornkamp and Ickstadt, 2009] Bornkamp, B. and Ickstadt, K. (2009). Bayesian nonparametric estimation of continuous monotone functions with applications to dose-response analysis. *Biometrics*, 65(1):198–205.

[Canini et al., 2016] Canini, K., Cotter, A., Gupta, M., Milani Fard, M., and Pfeifer, J. (2016). Fast and flexible monotonic functions with ensembles of lattices. In *Advances in Neural Information Processing Systems (NeurIPS)*.

[Chen et al., 2018] Chen, R., Rubanova, Y., Bettencourt, J., and Duvenaud, D. (2018). Neural ordinary differential equations. In *Advances in Neural Information Processing Systems (NeurIPS)*.

[Curtis and Ghosh, 2011] Curtis, S. M. and Ghosh, S. K. (2011). A variable selection approach to monotonic regression with bernstein polynomials. *Journal of Applied Statistics*, 38(5):961–976.

[Da Veiga and Marrel, 2012] Da Veiga, S. and Marrel, A. (2012). Gaussian process modeling with inequality constraints. *Annales de la Faculté des sciences de Toulouse : Mathématiques*, Ser. 6, 21(3):529–555.

[Dette and Scheder, 2006] Dette, H. and Scheder, R. (2006). Strictly monotone and smooth nonparametric regression for two or more variables. *Canadian Journal of Statistics*, 34(4):535–561.

[Durot and Lopuhaä, 2018] Durot, C. and Lopuhaä, H. (2018). Limit theory in monotone function estimation. *Statistical Science*, 33(4):547–567.

[Foong et al., 2019] Foong, A. Y. K., Burt, D. R., Li, Y., and Turner, R. E. (2019). Pathologies of factorised gaussian and mc dropout posteriors in bayesian neural networks.

[Golchi et al., 2015] Golchi, S., Bingham, D., Chipman, H., and Campbell, D. (2015). Monotone emulation of computer experiments. *SIAM-ASA Journal on Uncertainty Quantification*, 3(1):370–392.

[Hall and Huang, 2001] Hall, P. and Huang, L.-S. (2001). Nonparametric kernel regression subject to monotonicity constraints. *Annals of Statistics*, 29(3):624–647.

[Haslett and Parnell, 2008] Haslett, J. and Parnell, A. (2008). A simple monotone process with application to radiocarbon-dated depth chronologies. *Journal of the Royal Statistical Society. Series C*, 57:399–418.

[Hegde et al., 2019] Hegde, P., Heinonen, M., Lähdesmäki, H., and Kaski, S. (2019). Deep learning with differential gaussian process flows. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*.

[Heinonen et al., 2018] Heinonen, M., Yildiz, C., Mannerström, H., Intosalmi, J., and Lähdesmäki, H. (2018). Learning unknown ode models with gaussian processes. In *International Conference on Machine Learning (ICML)*.

[Kaiser et al., 2018] Kaiser, M., Otte, C., Runkler, T., and Ek, C. H. (2018). Bayesian alignments of warped multi-output gaussian processes. In *Advances in Neural Information Processing Systems (NeurIPS)*.

[Kazlauskaite et al., 2019] Kazlauskaite, I., Ek, C. H., and Campbell, N. (2019). Gaussian process latent variable alignment learning. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*. PMLR.

[Kim et al., 2018] Kim, D., Ryu, H., and Kim, Y. (2018). Nonparametric bayesian modeling for monotonicity in catch ratio. *Communications in Statistics: Simulation and Computation*, 47(4):1056–1065.

[Lavine and Mockus, 1995] Lavine, M. and Mockus, A. (1995). A nonparametric bayes method for isotonic regression. *Journal of Statistical Planning and Inference*, 46(2):235–248.

[Lawrence, 2004] Lawrence, N. D. (2004). Gaussian process latent variable models for visualisation of high dimensional data. *Advances in neural information processing systems*, 16(3):329–336.

[Lenk and Choi, 2017] Lenk, P. and Choi, T. (2017). Bayesian analysis of shape-restricted functions using gaussian process priors. *Statistica Sinica*, 27(1):43–69.

[Lin and Dunson, 2014] Lin, L. and Dunson, D. (2014). Bayesian monotone regression using gaussian process projection. *Biometrika*, 101(2):303–317.

[Lopez-Lopera et al., 2019] Lopez-Lopera, A. F., John, S., and Durrande, N. (2019). Gaussian process modulated cox processes under linear inequality constraints. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*.

[Lorenzi et al., 2019] Lorenzi, M., Filippone, M., Frisoni, G., Alexander, D., and Ourselin, S. (2019). Probabilistic disease progression modeling to characterize diagnostic uncertainty: Application to staging and prediction in alzheimer's disease. *NeuroImage*, 190:56–68.

[Maatouk, 2017] Maatouk, H. (2017). Finite-dimensional approximation of gaussian processes with inequality constraints. *arXiv:1706.02178*.

[Nader et al., 2019] Nader, C. A., Ayache, N., Robert, P., and Lorenzi, M. (2019). Monotonic gaussian process for spatio-temporal trajectory separation in brain imaging data. *arXiv:1902.10952*.

[Øksendal, 1992] Øksendal, B. (1992). *Stochastic Differential Equations (3rd Ed.): An Introduction with Applications*. Springer-Verlag.

[Raket et al., 2016] Raket, L. L., Grimme, B., Schöner, G., Igel, C., and Markussen, B. (2016). Separating timing, movement conditions and individual differences in the analysis of human movement. *PLOS Computational Biology*, 12(9):1–27.

[Ramsay, 1988] Ramsay, J. (1988). Monotone regression splines in action. *Statistical Science*, 3(4):425–441.

[Ramsay, 1998] Ramsay, J. (1998). Estimating smooth monotone functions. *Journal of the Royal Statistical Society. Series B*, 60(2):365–375.

[Rasmussen and Williams, 2005] Rasmussen, C. E. and Williams, C. K. I. (2005). *Gaussian Processes for Machine Learning*.

[Riihimäki and Vehtari, 2010] Riihimäki, J. and Vehtari, A. (2010). Gaussian processes with monotonicity information. *International Conference on Artificial Intelligence and Statistics (AISTATS)*.

[Shively et al., 2009] Shively, T. S., Sager, T. W., and Walker, S. G. (2009). A bayesian approach to non-parametric monotone function estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(1):159–175.

[Siivola et al., 2016] Siivola, E., Piironen, J., and Vehtari, A. (2016). Automatic monotonicity detection for gaussian processes. *arXiv:1610.05440*.

[Sill and Abu-Mostafa, 1997] Sill, J. and Abu-Mostafa, Y. (1997). Monotonicity hints. In *Advances in Neural Information Processing Systems (NeurIPS)*.

[Snoek et al., 2014] Snoek, J., Swersky, K., Zemel, R., and Adams, R. (2014). Input warping for bayesian optimization of non-stationary functions. In *International Conference on Machine Learning (ICML)*.

[Spengler, 1984] Spengler, E. (1984). Ghostbusters.

[Titsias, 2009] Titsias, M. (2009). Variational Learning of Inducing Variables in Sparse Gaussian Processes. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*.

[Ustyuzhaninov et al., 2019] Ustyuzhaninov, I., Kazlauskaite, I., Kaiser, M., Bodin, E., Campbell, N. D. F., and Ek, C. H. (2019). Compositional uncertainty in deep gaussian processes. In *Bayesian Deep Learning Workshop at NeurIPS*.

[Wahba, 1978] Wahba, G. (1978). Improper priors, spline smoothing and the problem of guarding against model errors in regression. *Journal of the Royal Statistical Society. Series B*, 49.

[Wolberg and Alfy, 2002] Wolberg, G. and Alfy, I. (2002). An energy-minimization framework for monotonic cubic spline interpolation. *Journal of Computational and Applied Mathematics*, 143(2):145–188.

[Yildiz et al., 2018a] Yildiz, C., Heinonen, M., Intosalmi, J., Mannerstrom, H., and Lahdesmaki, H. (2018a). Learning stochastic differential equations with gaussian processes without gradient matching. In *IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*.

[Yildiz et al., 2018b] Yildiz, C., Heinonen, M., and Lähdesmäki, H. (2018b). A nonparametric spatio-temporal SDE model. In *Spatiotemporal Workshop at NeurIPS*.