# Non-Gaussian Process Regression

**Yaman Kındap**
SigProC Laboratory
University of Cambridge
yk392@cam.ac.uk

**Simon Godsill**
SigProC Laboratory
University of Cambridge
sjg@eng.cam.ac.uk

## Abstract

Standard GPs offer a flexible modelling tool for well-behaved processes. However, deviations from Gaussianity are expected to appear in real world datasets, with structural outliers and shocks routinely observed. In these cases GPs can fail to model uncertainty adequately and may over-smooth inferences. Here we extend the GP framework into a new class of time-changed GPs that allow for straightforward modelling of heavy-tailed non-Gaussian behaviours, while retaining a tractable conditional GP structure through an infinite mixture of non-homogeneous GPs representation. The conditional GP structure is obtained by conditioning the observations on a latent transformed input space and the random evolution of the latent transformation is modelled using a Lévy process which allows Bayesian inference in both the posterior predictive density and the latent transformation function. We provide Markov chain Monte Carlo inference procedures for this model and demonstrate the potential benefits compared to a standard GP in detail in the appendix.

## 1 Introduction

The design of kernel functions that are able to represent a wide range of characteristics and make consistent generalisations is a fundamental area of research. Some recent work in this area include modelling the kernel via spectral densities that are scale-location mixtures of Gaussians [1], and similarly using Lévy process priors over adaptive basis expansions for the spectral density [2]. Spectral kernels are generalised to non-stationary kernels in [3]. For stationary time series models, a prior over nonparametric kernels can be defined through a separate GP [4].

Extensions to the standard GP model can be made by directly modelling the covariance matrix as a stochastic process [5], assuming heteroscedastic noise on the observations and carrying out variational inference [6], or learning nonlinear transformations of the observations such that the latent transformed observations are modelled well by a GP [7, 8]. Nonstationarity in the measurement process can be expressed as a product of multiple GPs [9] and heavy-tailed observations may be modelled through the Student-t process where the predictive kernel function depends on the values of the observations [10], unlike the standard GP where the kernel is determined only by the values of the input set. Particularly relevant extensions of GP models are presented in [11] where the input space is locally modelled by separate GPs, and string GPs [12] introduce link functions between local GPs such that the global process is still a GP and provides efficient inference methods on large data sets. In [13, 14] a latent space is defined between the inputs and observations through a separate GP and a class of bounded functions in $[0, 1]$, respectively. In [15] the inputs are assumed to be unobserved and integrated out using a variational approach which leads to deep GPs [16].

By designing expressive covariance functions or stacking multiple GPs in structured arrangements, the GP framework produces accurate predictive models in numerous application domains. However, these models are limited by their Gaussianity assumption such that the local patterns learned through these models are highly dependent on particular observations instead of learning the overall dynamics

2022 NeurIPS Workshop on Gaussian Processes, Spatiotemporal Modeling, and Decision-making Systems.

of the data generating system. A more natural and interpretable way to define complex relationships may be to assume that the underlying random function is non-Gaussian which yields more sparse representations [17].

In this work, we present a novel approach to modelling non-Gaussian dynamics by constructing a non-Gaussian process (NGP) such that the observations form a conditional GP that is conditioned on a latent input transformation function that is separately modelled as a Lévy process. Building on the definition of a stationary kernel, the latent layer between the input and output spaces represent the random distances between any two points on an input space. In order to define the distribution of random distances without referring to a specific origin, and in order to maintain monotonicity of the input space transformation, the latent space of transformation functions is modelled by a special class of Lévy process called a subordinator that is non-negative and non-decreasing. Such a process is characterised by the distribution of its stationary and independent increments which as a result defines a probability distribution over the distance between any two input values. Making random monotonic transformations of input values allow the kernel to adapt to the local characteristics of an input space or in other words to the varying rate of change in the observations and the learned subordinator provides uncertainty estimates over the variation of the observed process everywhere on its domain.

In this paper we focus principally upon 1-dimensional GPs for the sake of brevity, but we emphasise that our approach can be readily extended to multiple dimensions, as described throughout the text and illustrated in the experimental results.

NGPs are related to continuous-time stochastic volatility models studied in the mathematical finance literature to model the behaviour of a stochastic process which has a randomly distributed variance [18]. The time-change operation defined for continuous-time stochastic processes is a standard approach to building stochastic volatility models. A common example is the time-changed Brownian motion where the time-change is chosen to be a subordinator and the time-changed motion produces a Lévy process [19]. In such a model, the process is conditionally a Brownian motion i.e. the integral of a white-noise GP. Similarly, our construction of a stationary NGP follows a GP conditioned on the latent values of a subordinator, thus it is a time-changed GP. Particular non-Gaussian behaviour can be expressed through the characterisation of a subordinator, examples include the stable law, normal-tempered stable, and generalised hyperbolic (including Student-t) processes. Hence, NGPs provide a flexible and expressive probabilistic framework for nonparametric learning of functions.

## 2 Probabilistic model

Define a non-negative, non-decreasing stochastic process $\{W(t)\}_{t\geq 0}$ such that it randomly maps time instances while preserving their order, therefore changing the time. A time-changed stochastic process $\{f(t)\}_{t\geq 0}$ is then defined as $f(t) = g(W(t))$ where the evolution of $f$ is governed by $dW(t)$ instead of $dt$. In other words, the change in $f$ will have variance proportional to $W(t) - W(s)$, instead of $t - s$ where $t > s$. Assuming that $g(t)$ is Gaussian, this operation enables large deviations from Gaussian behaviour to occur when $dW(t)$ is large, while retaining a conditionally Gaussian form.

The random evolution of $W(t)$ can be modelled as a subordinator that take values in $[0, \infty)$ such that it has independent and stationary increments with no fixed discontinuities [20, 21]. Thus, a subordinator increases non-linearly with a certain statistical distribution defined by the random number of discontinuities and their random magnitudes. A Lévy process $W(t)$ in $[0, \infty)$ having no drift or Brownian motion is defined through its characteristic function $\mathbb{E}\left[\exp(iuW(t))\right] = \exp\left(t\left[\int_{(0,\infty)}(e^{iuw} - 1)Q(dw)\right]\right)$ ([22], Corollary 15.8) where $Q$ is a Lévy measure that satisfies $\int_{(0,\infty)}(1 \wedge x)Q(dx) < \infty$ ([21], p.72). The Lévy measure $Q$ is defined on the random magnitudes of discontinuities, called jumps, and denotes the expected number of jumps per unit time whose magnitudes belong to some subset of the jump space [23].

By the Lévy-Itô decomposition, a pure jump Lévy process (i.e. containing no Brownian motion) may be expressed using a stochastic integral as

$$W(t) = \int_{(0,\infty)} wN([0,t], dw) \tag{1}$$

where $N$ is a bivariate point process having mean measure $Leb \times Q$ on $[0, T] \times (0, \infty)$ which can be conveniently expressed using a Poisson random measure as

$$N = \sum_{i=1}^{\infty} \delta_{V_i, M_i} \tag{2}$$

where $\{V_i \in [0, T]\}$ are i.i.d. uniform random variables which give the times of arrival of jumps, $\{M_i\}$ are the sizes of the jumps and $\delta_{V_i, M_i}$ is Dirac measure centered at time $V_i$ and jump size $M_i$. Substituting $N$ into Eq. (1) leads to a representation of a Lévy jump process as an infinite series

$$W(t) = \sum_{i=1}^{\infty} M_i \mathcal{I}_{V_i \leq t} \quad a.s. \tag{3}$$

The almost sure convergence of this series to $\{W(t)\}$ is proved in [24]. Therefore, by sampling pairs of jump times and sizes $\{V_i, M_i\}$, a realisation of a Lévy process $W(t)$ may be obtained.

The standard formulation of the time-change operation on $[0, \infty)$ can be extended to $d$-dimensional input spaces $\mathcal{X}$ by considering the Poisson random measure representation $N$ of a Lévy process. A homogeneous Poisson process expressing the jump times can be generalised to any number of dimensions where we define arbitrary inputs $x', x \in \mathbb{R}^d$ [25]. The independence properties of a Lévy measure allow the definition on the unit time interval to be extended to unit $d$-dimensional volumes by appropriately scaling the rate of the process [26]. For multidimensional input transformations a subordination field on $\mathbb{R}^d$ is a $d$-dimensional stochastic process such that each of its dimensions is a subordinator. Thus the $i$-th dimension of an input vector $x^{(i)}$ is mapped to $W^{(i)}(x^{(i)})$ where $W^{(i)}$ denotes the subordinator on $i$. Therefore a distance $d(x', x)$ can be randomly transformed as $d(W(x'), W(x))$. Hence, the choice of a Lévy measure characterise the distribution of the random distances over the input space. The notation introduced for the multidimensional treatment of subordination is omitted for brevity in the following sections as it is straightforward to extend the model into multidimensional input spaces.

A non-Gaussian process (NGP) prior on functions can be obtained by randomly transforming the inputs using a subordinator and carrying out GP regression on the transformed input space. The resulting posterior distribution follows a non-Gaussian stochastic process. Given a set of input-output pairs $\{x_i, y_i\}$ consider a latent input transformation such that $x_i$ is mapped to $W(x_i)$ where $\{W(x); x \in \mathcal{X}\}$ is a subordinator. The associated prior on the transformation function is then defined as $p(W)$ and the conditional prior over $f$ is $p(f|W) \sim \mathcal{GP}(m_W(x), K_W(x', x))$ where $m_W(x) = m(W(x))$, $K_W(x', x) = K(W(x'), W(x)) = K(|W(x') - W(x)|)$ and $K(\cdot, \cdot)$ is a stationary kernel function e.g. squared exponential or Matérn. The joint distribution over the product space of $f$ and $W$, $p(f, W|y_{1:n})$ characterises the NGP prior.

The conditional GP structure of a NGP induces a posterior mean $\bar{m}_W(\cdot)$ and kernel function $\bar{K}_W(\cdot, \cdot)$ that can be evaluated analytically, i.e. $p(f|y_{1:n}, W) \sim \mathcal{GP}(\bar{m}_W, \bar{K}_W)$. The conditional likelihood $p(y_{1:n}|W)$ is of particular interest in this framework since it is a measure of how well the data is represented by the model given a random transformation and it can also be evaluated analytically.

The NGP posterior distribution over the function space is found as

$$p(f|y_{1:n}) = \int p(f|y_{1:n}, W) p(W|y_{1:n}) dW$$

where $p(W|y_{1:n})$ is the posterior distribution of the subordinator process. Inferring $p(W|y_{1:n})$ and hence $p(f|y_{1:n})$ is analytically intractable, however using approximate inference methods allow for straightforward extensions of the model and fully Bayesian inference.

Our work relies on shot-noise simulation methods for simulating Lévy processes based on series representations [24] and we describe a novel Metropolis-Hastings-within-Gibbs (MH-in-Gibbs) algorithm [27, 28] to obtain samples from the posterior distribution of a subordinator and estimate a non-Gaussian process posterior $p(f|y_{1:n})$ in the appendix.

# 3 Experimental Results and Discussion

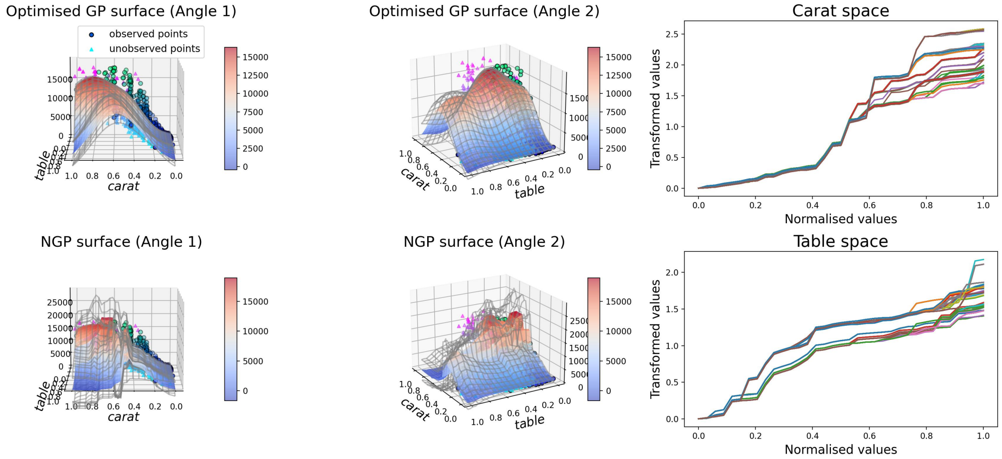

Figure 1: Regression analysis results for NGP and GP models with for the diamond price data set using a TS subordinator. The posterior means are plotted as a surface and the ±3 standard deviation surface is overlaid on the mean as a wireframe plot. The right hand column shows the posterior subordinator samples for a TS subordinator.

In Fig. 1 we present the results of applying NGP regression to a non-Gaussian two dimensional data set available in TensorFlow [29] on diamond prices and compare the results with an alternative standard GP regression setting. The features used in this task are the carat of a diamond which is a measure of its weight and the percentage length of its table which is the largest flat facet of the diamond and affects how the diamond interacts with light. For ease of visualisation, both input dimensions are linearly transformed to lie between $[0, 1]$. The experiment is designed such that a $1000$ randomly selected input-output pairs are chosen as the training set and the learned posterior surface is compared against another randomly selected $1000$ pairs.

The main distinguishing property of the diamond price data set is that the price increases non-linearly with increases in the carat feature. The GP surface shown here forms a smooth function that cannot model the rapid change in prices around these regions of non-linear increase. Furthermore, the predictive surface is making non-zero predictions around out-of-sample regions which are undesirable in downstream decision-making tasks such as in Bayesian optimisation problems. Alternatively, the NGP surface is able to identify the non-linear increase in prices and the increased uncertainty in prices for larger carat values. These properties can be most clearly identified in the posterior subordinator samples. Note that the mean log conditional likelihood of the MCMC samples are found as $-97873.9$ and the GP log likelihood is found to be $-113731.8$.

NGP regression with a tempered stable subordinator presented in this work may be applied to datasets where there are local deviations from Gaussian behaviour but the overall trend of the function closely follows a GP. Using alternative characterisations of the subordinator, varying degrees of non-Gaussian behaviour can be modelled in a NGP regression framework. Using NGP regression produces a generative model conditioned on a dataset where samples from both the posterior density over the input transformation and the predictive density over the observations can be generated. A probabilistic representation of a latent layer between the input and output spaces may lead to new insights about the underlying data generating mechanism. Furthermore, our construction of a NGP using the time-change operation may potentially be extended to any probabilistic setting for interpreting non-Gaussian behaviour. For example, Lévy fields can be used to model the first layer of a deep GP architecture ([15, 16]).

## References

[1] A. Wilson and R. Adams, "Gaussian Process Kernels for Pattern Discovery and Extrapolation," in *Proceedings of the 30th International Conference on Machine Learning* (S. Dasgupta and D. McAllester, eds.), vol. 28 of *Proceedings of Machine Learning Research*, (Atlanta, Georgia, USA), pp. 1067–1075, PMLR, 17–19 Jun 2013.

[2] P. A. Jang, A. Loeb, M. Davidow, and A. G. Wilson, "Scalable Levy Process Priors for Spectral Kernel Learning," in *Advances in Neural Information Processing Systems* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), vol. 30, Curran Associates, Inc., 2017.

[3] Y.-L. K. Samo and S. Roberts, "Generalized spectral kernels," *arXiv preprint arXiv:1506.02236*, 2015.

[4] F. Tobar, T. D. Bui, and R. E. Turner, "Learning Stationary Time Series using Gaussian Processes with Nonparametric Kernels," in *Advances in Neural Information Processing Systems* (C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, eds.), vol. 28, Curran Associates, Inc., 2015.

[5] A. G. Wilson and Z. Ghahramani, "Generalised Wishart Processes," 2011.

[6] M. Lázaro-Gredilla and M. Titsias, "Variational Heteroscedastic Gaussian Process Regression," pp. 841–848, 01 2011.

[7] E. Snelson, Z. Ghahramani, and C. Rasmussen, "Warped Gaussian Processes," in *Advances in Neural Information Processing Systems* (S. Thrun, L. Saul, and B. Schölkopf, eds.), vol. 16, MIT Press, 2003.

[8] M. Lázaro-Gredilla, "Bayesian Warped Gaussian Processes," in *Advances in Neural Information Processing Systems* (F. Pereira, C. Burges, L. Bottou, and K. Weinberger, eds.), vol. 25, Curran Associates, Inc., 2012.

[9] R. P. Adams and O. Stegle, "Gaussian Process Product Models for Nonparametric Nonstationarity," in *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, (New York, NY, USA), p. 1–8, Association for Computing Machinery, 2008.

[10] A. Shah, A. Wilson, and Z. Ghahramani, "Student-t Processes as Alternatives to Gaussian Processes," in *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics* (S. Kaski and J. Corander, eds.), vol. 33 of *Proceedings of Machine Learning Research*, (Reykjavik, Iceland), pp. 877–885, PMLR, 22–25 Apr 2014.

[11] C. Rasmussen and Z. Ghahramani, "Infinite Mixtures of Gaussian Process Experts," in *Advances in Neural Information Processing Systems* (T. Dietterich, S. Becker, and Z. Ghahramani, eds.), vol. 14, MIT Press, 2001.

[12] Y.-L. Samo and S. J. Roberts, "String and membrane Gaussian processes," *Journal of Machine Learning Research*, vol. 17, 2016.

[13] A. M. Schmidt and A. O'Hagan, "Bayesian Inference for Non-Stationary Spatial Covariance Structure via Spatial Deformations," *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, vol. 65, no. 3, pp. 743–758, 2003.

[14] J. Snoek, K. Swersky, R. Zemel, and R. Adams, "Input Warping for Bayesian Optimization of Non-Stationary Functions," in *Proceedings of the 31st International Conference on Machine Learning* (E. P. Xing and T. Jebara, eds.), vol. 32 of *Proceedings of Machine Learning Research*, (Bejing, China), pp. 1674–1682, PMLR, 22–24 Jun 2014.

[15] M. Titsias and N. D. Lawrence, "Bayesian Gaussian Process Latent Variable Model," in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics* (Y. W. Teh and M. Titterington, eds.), vol. 9 of *Proceedings of Machine Learning Research*, (Chia Laguna Resort, Sardinia, Italy), pp. 844–851, PMLR, 13–15 May 2010.

[16] A. Damianou and N. D. Lawrence, "Deep Gaussian processes," in *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics* (C. M. Carvalho and P. Ravikumar, eds.), vol. 31 of *Proceedings of Machine Learning Research*, (Scottsdale, Arizona, USA), pp. 207–215, PMLR, 29 Apr–01 May 2013.

[17] M. Unser and P. D. Tafti, *An Introduction to Sparse Stochastic Processes.* Cambridge University Press, 2014.

[18] E. Ghysels, A. C. Harvey, and E. Renault, "Stochastic volatility," in *Statistical Methods in Finance*, vol. 14 of *Handbook of Statistics*, pp. 119–191, Elsevier, 1996.

[19] A. Veraart and M. Winkel, "Time change," in *Encyclopedia of Quantitative Finance*, pp. 1812–1816, Wiley, 2010.

[20] W. Feller, *An introduction to probability theory and its applications*. No. v. 2 in Wiley mathematical statistics series, Wiley, 1966.

[21] J. Bertoin, *Lévy Processes*. Cambridge Tracts in Mathematics, 121, Cambridge University Press, 1997.

[22] O. Kallenberg, *Foundations of Modern Probability*. Springer-Verlag, 2nd ed., 2002.

[23] Tankov, P. and Cont, R., *Financial Modelling with Jump Processes, Second Edition*. Chapman and Hall/CRC Financial Mathematics Series, Taylor & Francis, 2015.

[24] J. Rosiński, *Series Representations of Lévy Processes from the Perspective of Point Processes*, pp. 401–415. Boston, MA: Birkhäuser Boston, 2001.

[25] Kingman, J.F.C., *Poisson Processes*. Oxford Studies in Probability, Clarendon Press, 1992.

[26] R. L. Wolpert and K. Ickstadt, *Simulation of Lévy Random Fields*, pp. 227–242. New York, NY: Springer New York, 1998.

[27] W. K. Hastings, "Monte Carlo sampling methods using Markov chains and their applications," *Biometrika*, vol. 57, no. 1, pp. 97–109, 1970.

[28] S. Chib and E. Greenberg, "Understanding the Metropolis-Hastings Algorithm," *The American Statistician*, vol. 49, no. 4, pp. 327–335, 1995.

[29] H. Wickham, *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016.

[30] T. S. Ferguson and M. J. Klass, "A Representation of Independent Increment Processes without Gaussian Components," *The Annals of Mathematical Statistics*, vol. 43, no. 5, pp. 1634 – 1643, 1972.

[31] R. Wolpert and K. Ickstadt, "Poisson/gamma random field models for spatial statistics," *Biometrika*, vol. 85, pp. 251–267, 06 1998.

[32] S. Godsill and Y. Kındap, "Point process simulation of generalised inverse Gaussian processes and estimation of the Jaeger integral," *Statistics and Computing*, vol. 32, p. 13, Dec 2021.

[33] P. A. W. Lewis and G. S. Shedler, "Simulation of nonhomogeneous Poisson processes by thinning," *Naval Research Logistics Quarterly*, vol. 26, pp. 403–413, September 1979.

[34] P. Carr, H. Geman, D. B. Madan, and M. Yor, "Stochastic Volatility for Lévy Processes," *Mathematical Finance*, vol. 13, no. 3, pp. 345–382, 2003.

[35] S. Godsill, M. Riabiz, and I. Kontoyiannis, "The Lévy State Space Model," in *2019 53rd Asilomar Conference on Signals, Systems, and Computers*, pp. 487–494, 2019.

[36] J. Rosiński, "Tempering stable processes," *Stochastic Processes and their Applications*, vol. 117, no. 6, pp. 677–707, 2007.

[37] N. Shephard and O. E. Barndorff-Nielsen, "Normal Modified Stable Processes," Economics Series Working Papers 72, University of Oxford, Department of Economics, July 2001.

[38] J. Imai and R. Kawai, "On finite truncation of infinite shot noise series representation of tempered stable laws," *Physica A-statistical Mechanics and Its Applications*, vol. 390, pp. 4411–4425, 2011.

[39] G. Samorodnitsky and M. S. Taqqu, *Stable non-Gaussian random processes : stochastic models with infinite variance*. CRC Press, 1994.

[40] A. E. Gelfand, "Gibbs sampling," *Journal of the American statistical Association*, vol. 95, no. 452, pp. 1300–1304, 2000.

[41] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*. MIT Press, 2006.

[42] O. Barndorff-Nielsen, J. Kent, and M. Sørensen, "Normal Variance-Mean Mixtures and z Distributions," *International Statistical Review / Revue Internationale de Statistique*, vol. 50, no. 2, pp. 145–159, 1982.

[43] W. P. Bruinsma, M. Tegnér, and R. E. Turner, "Modelling Non-Smooth Signals With Complex Spectral Structure," in *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, Proceedings of Machine Learning Research, PMLR, 2022.

# A Sampling and Inference

In this section, we review shot-noise simulation methods for simulating Lévy processes based on series representations [24]. We describe a novel Metropolis-Hastings-within-Gibbs (MH-in-Gibbs) algorithm [27, 28] to obtain samples from the posterior distribution of a subordinator and estimate a non-Gaussian process posterior $p(f|y_{1:n})$.

## A.1 Shot-noise simulation methods

The jump magnitudes $\{M_i\}_{i=1}^{\infty}$ shown in Eq. (3) of a Lévy process cannot be directly simulated because there may be an infinite number of jumps in any finite interval. One way to obtain approximate samples from such an infinite sequence is to consider ordering the jump magnitudes by size and simulating large jumps while ignoring or approximating the residual error as discussed in [30, 24, 31, 26, 32]. Once the ordered jump sizes have been obtained, the corresponding jump positions $\{V_i\}_{i=1}^{\infty}$ may be simulated independently from a uniform distribution on $(x_{lb}, x_{ub})$ where $x_{lb}$, $x_{ub}$ are some lower and upper bounds, or sequentially in space from a homogeneous Poisson process if preferred.

Consider a bivariate point process $N'$ that has the same form as Eq. (2) where the jump magnitudes $M_i$ are expressed as the output of a function $h(\Gamma_i)$ where $\{\Gamma_i\}_{i=1}^{\infty}$ are the epochs of a unit rate Poisson process, i.e. the cumulative sum of exponential random variables with unit rate, independent of $\{V_i\}_{i=1}^{\infty}$. Similar to the standard inverse CDF method for random variate generation, the upper tail mass of a Lévy measure $Q^+(x) = Q([x, \infty))$ can be inverted to produce jump magnitudes of a subordinator by passing epochs of a homogeneous Poisson process through the inverse Lévy measure $Q^{+^{-1}}(\cdot)$. The corresponding function $h(\cdot) = Q^{+^{-1}}(\cdot)$ is non-increasing thus $\{h(\Gamma_i)\}$ is an ordered sequence representing random jump sizes. Note that the epochs of a homogeneous Poisson process are analogous to uniformly distributed random variables in $(0, \infty)$ and the mapping theorem states that the resulting points $\{V_i, h(\Gamma_i)\}$ form a Poisson point process $N' = \sum_{i=1}^{\infty} \delta_{V_i, h(\Gamma_i)}$ on $(x_{lb}, x_{ub}) \times (0, \infty)$ [25]. $N'$ converges almost surely to $N$ as the $\{\Gamma_i\}$ sequence is simulated indefinitely [24] and approximations of the point process may be obtained through finite samples.

The explicit evaluation of the inverse tail measure $Q^{+^{-1}}(\cdot)$ in general is not possible. The Lévy measures considered in this paper possess a density function denoted as $Q(x)$ such that $Q(dx) = Q(x)dx$. The approach taken in this work is to simulate from a tractable dominating point process $N_0$ having Lévy measure $Q_0$ such that $Q_0(dx)/Q(dx) \geq 1,\ \forall x \in (0, \infty)$ for which $h(\cdot)$ can be explicitly evaluated. The resulting jump magnitudes belonging to $N_0$ are then thinned with probability $Q(x)/Q_0(x)$ as in [33] to obtain the desired approximate jump magnitudes $\{M_i\}$ of a subordinator.

As a motivating example in this paper, we consider tempered stable (TS) processes which are commonly used in mathematical finance to model stochastic volatility [34]. We note that our methodology is equally applicable to other subordinator processes for which shot noise simulation methods can be applied [32, 35]. A TS process exhibits both $\alpha$-stable and Gaussian trends depending on the distance it travels. For short distances the stable characteristics prevail and the TS process produces larger jumps compared to a Gaussian process. For longer distances the tempering causes a TS process to produce Gaussian trends [36]. Thus, a TS process is a natural extension to Gaussian processes.

The Lévy density for the subordinator TS process is defined as [37, 36]

$$Q(x) = Cx^{-1-\alpha}e^{-\beta x}, \qquad x > 0 \tag{4}$$

where $\alpha \in (0, 1)$ is the tail parameter and $\beta$ is the tempering parameter. The corresponding tail probability may be calculated in terms of gamma functions but it cannot be analytically inverted and numerical approximations are needed [38]. Instead, we adopt a thinning approach where the Lévy density is factorised into a $\alpha$-stable subordinator process with Lévy density $Q_0(x) = Cx^{-1-\alpha}$ [39, 35] and a tempering function $e^{-\beta x}$. The tail mass of a stable process can be found to be

$Q_0^+(x) = \frac{C}{\alpha}x^{-\alpha}$ and inverting this function produces the simulation function $h(\gamma) = \left(\frac{\alpha\gamma}{C}\right)^{-1/\alpha}$. Given points $x_i$ from a stable point process with density $Q_0(x)$, individually selecting (thinning) points with probability $e^{-\beta x_i}$ results in a tempered stable process. The associated sampling algorithm is shown in Alg. 1 for reference.

**Algorithm 1** Generation of the jumps of a tempered stable process with Lévy density $Q_{TS}(x) = Cx^{-1-\alpha}e^{-\beta x}$ where $\alpha$ is the tail parameter and $\beta$ is the tempering parameter.

```
1. Assign $N_{TS} = \emptyset$,
2. Generate the epochs of a unit rate Poisson process, $\{\Gamma_i;\ i = 1, 2, 3...\}$,
3. For $i = 1, 2, 3...$,
     • Compute $x_i = \left(\frac{\alpha\Gamma_i}{C}\right)^{-1/\alpha}$,
     • With probability $e^{-\beta x_i}$, accept $x_i$ and assign $N_{TS} = N_{TS} \cup x_i$.
```

Algorithm 1 generates the jumps that correspond to a TS process in $(0, 1)$. Since the jumps of a Lévy process are independent and stationary it is straightforward to adjust the interval. For instance, setting the rate of the underlying Poisson process produced in the second stage of Alg. 1 to the length of the interval $(x_{lb}, x_{ub})$ produces the associated TS process. Similarly, for $d$-dimensional input spaces the jumps on a $n$-dimensional hypercube can be simulated by setting the rate to the associated volume [26].

### A.2 Approximate inference

Since a stochastic process is defined as an infinite collection of random variables, designing direct sampling methods from the posterior $p(W|y_{1:n})$ based on batch Monte Carlo methods is a difficult task. Instead a more appropriate approach to high dimensional problems is to use a Gibbs sampler which approximates samples from a multivariate probability distribution or in this case a stochastic process. The latent random variables are grouped into smaller and more manageable collections, then each collection is iteratively updated conditioned on the previous samples and observations. The sequence of samples from such an algorithm can be considered as a Markov chain where the stationary distribution is the high dimensional posterior distribution that was targeted. For a short tutorial on Gibbs sampling see [40].

A Gibbs sampler approximating samples from $p(W|y_{1:n})$ can be implemented by simulating the associated bivariate random points that define the jump size and position on small disjoint intervals $\tau = (x_j, x_l)$ conditioned on the previous sample points in $-\tau = \mathcal{X} \setminus (x_j, x_l)$ and observations. Progressively simulating these points such that the whole input space is covered leads to approximate samples from the target distribution. Let $\{V_i^{(k)}, M_i^{(k)}\}$ be a random length sequence of jump positions and magnitudes associated with the $k$-th sample $W^{(k)}$ in a Monte Carlo procedure. For any interval $(x_j, x_l)$ new jump position and magnitudes $\{V_i^{(\prime)}, M_i^{(\prime)}\}$ can be simulated with a rate determined by the distance $|x_j - x_l|$ while removing the points associated with the same interval from $W^{(k)}$. The resulting sample path is denoted as $W^{(\prime)}$ before accepting or rejecting it as the $k+1$-th sample $W^{(k+1)}$.

**Algorithm 2** Simulating sample paths from the proposal density $p(W_\tau|W_{-\tau})$.

```
Given a random length set $N_W = \{V_i^{(k)}, M_i^{(k)}\}$ and an interval $(x_j, x_l) \in \mathcal{X}$,
1. Simulate $\{V_i^{(\prime)}, M_i^{(\prime)}\}$ with rate $|x_j - x_l|$ using Algorithm 1,
2. Remove all points $\{V_i^{(k)}, M_i^{(k)}\}$ from $N_W$ such that $x_j < V_i^{(k)} < x_l$ and add $\{V_i^{(\prime)}, M_i^{(\prime)}\}$,
   $N_W = N_W \cup \{V_i^{(\prime)}, M_i^{(\prime)}\}$,
3. Substitute the points of $N_W$ into Eq. (3) to obtain the proposed sample path $W^{(\prime)}$.
```

While Gibbs sampling reduces the complexity of sampling a stochastic process for each small interval, direct sampling from the conditional posterior for each interval is still intractable in general. Thus for each interval a Metropolis-Hastings algorithm is used yielding a MH-within-Gibbs sampling algorithm [28]. The proposal density for the MCMC sampling procedure is $p(W_\tau|W_{-\tau})$ which produces new bivariate points (jump sizes and times) on some interval $\tau$ conditioned on all points in $-\tau$ as described in Alg. 2.

For each realisation of the subordinator $W^{(k)}$, the conditional likelihood $p(y_{1:n}|W^{(k)})$ may be used as a weight in a Markov chain Monte Carlo sampler since it is proportional to the posterior distribution $p(W|y_{1:n})$ and we are proposing from $p(W_\tau|W_{-\tau})$. The conditional likelihood $p(y_{1:n}|W^{(k)})$ may be analytically found given the values of $W^{(k)}$. Then given a sample $W^{(k)}$ and proposal $W^{(\prime)}$, the acceptance probability for the proposal is

$$\alpha(W^{(\prime)}, W^{(k)}) = \text{min}\left(1, \frac{p(y_{1:n}|W^{(\prime)})}{p(y_{1:n}|W^{(k)})}\right) \tag{5}$$

**Algorithm 3** MH-within-Gibbs sampler for $p(W|y_{1:n})$.

```
1. Initialise $W^{(0)}$ by simulating $\{V_i, M_i\}$ from the associated bivariate point process using Alg. 1,
2. Analytically evaluate $\bar{m}_{W^{(0)}}$, $\bar{K}_{W^{(0)}}$ which define the conditional GP posterior $p(f|y_{1:n}, W^{(0)})$, and the conditional likelihood $p(y_{1:n}|W^{(0)})$,
3. For $N$ times, iterate over $\tau_j \in \mathcal{X}$ where $\bigcup_{j=1}^{J} \tau_j = \mathcal{X}$,
    (a) Using $\tau_j$ and the points $\{V_i^{(k)}, M_i^{(k)}\}$ associated with $W^{(k)}$, sample a proposed sample path $W^{(\prime)}$ using Alg. 2,
    (b) Evaluate $\bar{m}_{W^{(\prime)}}$, $\bar{K}_{W^{(\prime)}}$ and $p(y_{1:n}|W^{(\prime)})$,
    (c) With probability $\alpha(W^{(\prime)}, W^{(k)})$ the proposal is accepted and $W^{(k+1)} = W^{(\prime)}$, otherwise reject and set $W^{(k+1)} = W^{(k)}$.
```

The MH-within-Gibbs sampling procedure is described in Alg. 3. The resulting samples $\{W^{(k)}\}$ are individually associated with conditional GP posterior functions $p(f|y_{1:n}, W^{(k)})$ that are completely defined through their mean $\bar{m}_{W^{(k)}}$ and covariance $\bar{K}_{W^{(k)}}$ functions. Such a collection forms a Gaussian mixture distribution and the mean and covariance of the corresponding mixture density can be obtained as

$$\mathbb{E}_{f|\mathbf{y}}[f] = \frac{1}{N}\sum_{k=1}^{N} \bar{m}_{W^{(k)}} = m_{f|\mathbf{y}} \tag{6}$$

and

$$\text{Cov}_{f|\mathbf{y}}(f) = \frac{1}{N}\sum_{k=1}^{N} \left[\bar{K}_{W^{(k)}} + (\bar{m}_{W^{(k)}} - m_{f|\mathbf{y}})(\bar{m}_{W^{(k)}} - m_{f|\mathbf{y}})^T\right] \tag{7}$$

where $N$ is the number of samples and $\mathbb{E}_{f|\mathbf{y}}[f]$, $\text{Cov}_{f|\mathbf{y}}(f)$ define the posterior mean and covariance of the random function $f$. It is straightforward to obtain the corresponding predictive density $p(y^*|y_{1:N})$ by adding the observation noise matrix $\mathbf{\Omega}$ to each covariance matrix sample $\bar{K}_{W^{(k)}}$. Using a constant noise matrix $\mathbf{\Omega}$ corresponds to the assumption that the observation likelihood model is Gaussian [41]. This assumption can be relaxed by sampling a noise matrix $\mathbf{\Omega}^{(k)}$ for each individual sample to consider non-Gaussian likelihood models such as scale mixture of normals which includes the Student-t and Laplace distributions [42]. This results in doubly non-Gaussian behaviour which is highly expressive while retaining interpretation of individual components of the behaviour.

Following a similar approach the hyperparameters $C$, $\alpha$ and $\beta$ of the subordinator process may be included in the sampling procedure by considering an appropriate prior distribution over their values.

Hence these parameters may be marginalised out using the Monte Carlo procedure, which leaves the same number of kernel parameters that define a standard GP. This approach works successfully and will be reported in a future publication. Furthermore, a nonparametric kernel may be included in this framework by considering a prior distribution on stationary kernel functions and sampling a kernel function for each proposed sample. Some examples of nonparametric kernel design can be found in [1, 4, 43].

A straightforward extension of Alg. 3 to multidimensional input spaces can be achieved by assuming that individual subordinator dimensions $x^{(i)}$ are independent *a priori*. The simulation steps defined by Alg. 1 and 2 can be independently applied to each dimension and the other steps remain unchanged, replacing step 3. (b) with the multidimensional GP likelihood.