# 11 Gaussian and Dirichlet Processes

We encountered stochastic processes when we discussed noise in gravitational wave detectors and then again in the discussion of Time Series. Another application of stochastic processes is to generate probability distributions, as the relative frequencies of different outcomes of the stochastic process over long time intervals. We will be concerned with two particular types of stochastic process.

- **Gaussian processes:** These are infinite dimensional generalisations of the Normal distribution and realisations of these are random fields.
- **Dirichlet processes:** These are infinite dimensional generalisations of the Dirichlet distribution, and realisations of these are probability distributions.

## 11.1 Gaussian processes

A multivariate Gaussian distribution returns values of a finite set of random variables. A natural extension is to regard the set of random variables as the values of some random field at certain points. To generate the full random field we need an infinite dimensional Gaussian distribution, which is a Gaussian process. Formally we denote a random field, $y(\mathbf{x})$, generated by a Gaussian process via

$$y(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}'))$$

where $m(\mathbf{x})$ and $k(\mathbf{x}, \mathbf{x}')$ are the mean and covariance function of the Gaussian process. For simplicity of notation we assume that the random field is single valued at each point, but the extension to multivariate outputs is straightforward.

Formally, a GP is an infinite collection of variables, any finite subset of which are distributed as a multivariate Gaussian. For a set of parameter points $\{\mathbf{x}_i\}$, including, but not limited to, the training set $\mathcal{D}$,

$$[y(\mathbf{x}_i)] \sim N(\boldsymbol{m}, \boldsymbol{K}) , \tag{127}$$

where the mean vector and covariance matrix of this Gaussian distribution are fixed by the corresponding functions of the GP,

$$[\boldsymbol{m}]_i = m(\mathbf{x}_i) , \quad [\boldsymbol{K}]_{ij} = k(\mathbf{x}_i, \mathbf{x}_j) , \tag{128}$$

with probability density function

$$P\left(\{y(\mathbf{x}_i\}\right) = \frac{1}{\sqrt{(2\pi)^N |\boldsymbol{K}|}} \exp\left(-\frac{1}{2}\sum_{i,j}(y(\mathbf{x}_i) - m(\mathbf{x}_i)) \left[\boldsymbol{K}^{-1}\right]_{ij} (y(\mathbf{x}_j) - m(\mathbf{x}_j))\right) . \tag{129}$$

Gaussian processes are often used for interpolation. In that context, the training set $\mathcal{D}$ represents the set of known values of the field, e.g., the results of computational simulations at certain choices of input parameters, which we denote by $\tilde{y}(\mathbf{x}_i)$. The Gaussian process is constrained by this training set and then used to predict the value of the field at new points in the parameter space, with associated uncertainties. If the values of the field at the training points are not known perfectly, but have uncertainties $\epsilon_i \sim N(0, \sigma_i^2)$, the expression above takes the same form but with the replacement

$$[\boldsymbol{K}]_{ij} = k(\mathbf{x}_i, \mathbf{x}_j) \, + \sigma_i^2 \delta_{ij}.$$

Even with perfect simulations it can be advantageous to include a small error term, as this helps with inversion of the covariance matrix.

The mean and variance of the GP determine how the function is interpolated across the parameter space. It is common in regression to set the mean of the Gaussian process to zero, but specifying the covariance function is central to GP regression as it encodes our prior expectations about the properties of the function being interpolated. Possibly the simplest and most widely used choice for the covariance function is the squared exponential (SE)

$$k(\mathbf{x}_i, \mathbf{x}_j) = \sigma_f^2 \exp\left[-\frac{1}{2} g_{ab} (\mathbf{x}_i - \mathbf{x}_j)^a (\mathbf{x}_i - \mathbf{x}_j)^b\right] , \tag{130}$$

which defines a stationary, smooth GP. In Eq. (130), a scale $\sigma_f$ and a (constant) metric $g_{ab}$ for defining a modulus in parameter space have been defined. These are called hyperparameters and we denote them as $\vec{\theta} = \{\sigma_f, g_{ab}\}$, with Greek indices $\mu$, $\nu$, ... to label the components of this vector.

The probability in Eq. (129) is referred to as the hyperlikelihood, or alternatively the evidence for the training set; it is the probability that that particular realisation of waveform differences was obtained from a GP with a zero mean and specified covariance function. The hyperlikelihood depends only on the hyperparameters and the quantities in the training set, so we denote it as $Z(\vec{\theta}|\mathcal{D})$. The log hyperlikelihood is

$$\begin{aligned} \ln Z(\vec{\theta}|\mathcal{D}) \;\; = \;\; & -\frac{N}{2}\ln(2\pi) \\ & -\frac{1}{2}\sum_{i,j} (y(\mathbf{x}_i) - m(\mathbf{x}_i)) \left[k(\mathbf{x}_i, \mathbf{x}_j)\right]^{-1} (y(\mathbf{x}_j) - m(\mathbf{x}_j)) \\ & -\frac{1}{2}\ln\left|\det\left[k(\mathbf{x}_i, \mathbf{x}_j)\right]\right| . \end{aligned} \tag{131}$$

The values of the hyperparameters can be fixed to their optimum values $\vec{\theta}_{\rm op}$, defined as those which maximise the hyperlikelihood:

$$\left.\frac{\partial Z(\vec{\theta}|\mathcal{D})}{\partial \theta^\mu}\right|_{\vec{\theta}=\vec{\theta}_{\rm op}} = 0 \, . \tag{132}$$

An alternative approach is to consider the hyperparameters as nuisance parameters in addition to the source parameters $\mathbf{x}$, and marginalise over them while sampling an expanded likelihood,

$$\Lambda_{\rm expanded}(\mathbf{x}, \vec{\theta}|\mathcal{D}) \;\; \propto \mathcal{L}(\mathbf{x}|\vec{\theta}, \mathcal{D}) Z(\vec{\theta}|\mathcal{D}). \tag{133}$$

The disadvantage of this approach is that the hyperlikelihood is expensive to compute and the inclusion of extra nuisance parameters slows down any application of the GP. In contrast, maximising the likelihood is a convenient heuristic which is widely used in other contexts and allows all the additional computation to be done offline.

Having fixed the properties of the covariance function by examining the training set, we can now move on to using the GP as a predictive tool. The defining property of the GP is that any finite collection of variables drawn from it is distributed as a multivariate Gaussian in the manner of Eq. (129). Therefore, the set of variables formed by the training set plus the field at a set of extra parameter points $\{y(\mathbf{z}_j)\}$ is distributed as

$$\begin{bmatrix} y(\mathbf{x}_i) \\ y(\mathbf{z}_j) \end{bmatrix} \sim N\left(\boldsymbol{m}, \boldsymbol{\Sigma}\right) , \quad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{K} & \boldsymbol{K}_* \\ \boldsymbol{K}_*^{\mathrm{T}} & \boldsymbol{K}_{**} \end{pmatrix} , \tag{134}$$

where $\boldsymbol{K}$ is defined in Eq. (128) and the matrices $\boldsymbol{K}_*$ and $\boldsymbol{K}_{**}$ are defined as

$$[\boldsymbol{K}_*]_{ij} = k(\mathbf{x}_i, \mathbf{z}_j) , \quad [\boldsymbol{K}_{**}]_{ij} = k(\mathbf{z}_i, \mathbf{z}_j) . \tag{135}$$

The conditional distribution of the unknown field values at the new points, given the observed values in $\mathcal{D}$, can now be found and is given by

$$p(\{y(\mathbf{z}_i)\}) \propto \exp\left[ -\frac{1}{2} \sum_{j,k} (y(\mathbf{z}_j) - \mu_j) \Sigma^{-1}_{jk} (y(\mathbf{x}_k) - \mu_k) \right] \tag{136}$$

where the GPR mean and its associated error are given by

$$\mu_i = m(\mathbf{z}_i) + \sum_{j,k} [\boldsymbol{K}_*]_{ji} \left[\boldsymbol{K}^{-1}\right]_{jk} (\tilde{y}(\mathbf{x}_k) - m(\mathbf{x}_k)) , \tag{137}$$

$$\Sigma_{ij} = [\boldsymbol{K}_{**}]_{ij} - \sum_{k,l} [\boldsymbol{K}_*]_{ki} \left[\boldsymbol{K}^{-1}\right]_{kl} [\boldsymbol{K}_*]_{lj} . \tag{138}$$

## 11.2 The covariance function

The properties of the covariance function play an important role in determining the nature of the Gaussian process and its behaviour when used for regression. The only necessary requirements we have of a covariance function are that it is a positive definite; i.e. for any choice of points $\{\mathbf{x}_i\}$ the covariance matrix $K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$ is positive definite. The covariance function (and the corresponding GP) is said to be stationary if the covariance is a function only of $\vec{\tau} = \mathbf{x}_1 - \mathbf{x}_2$, furthermore it is said to be isotropic if it is a function only of $\tau \equiv |\vec{\tau}| = |\mathbf{x}_1 - \mathbf{x}_2|$.[3] Isotropy of a GP implies stationarity, but the converse is not true.

An example of how the properties of the covariance function relate to the properties of the GP, and hence the properties of the resulting interpolant, is given by considering the mean-square (MS) continuity and differentiability of GPs. It can be shown that the first $\zeta$ MS derivatives of a GP are MS continuous (the GP is said to be $\zeta$-times MS differentiable) if and only if the first $2\zeta$ derivatives of the covariance function are continuous at the diagonal point $\mathbf{x}_1 = \mathbf{x}_2 = \mathbf{x}_*$. For a stationary GP this condition reduces to checking the $2\zeta$ derivatives of $k(\vec{\tau})$ at $\vec{\tau} = \vec{0}$, and for an isotropic GP checking the $2\zeta$ derivatives of $k(\tau)$ at $\tau = 0$.

It is the smoothness properties of the covariance function at the origin that determine the differentiability of the GP. In the following subsections, we consider two aspects that enter the definition of the covariance function:

1. specifying the distance metric in parameter space $g_{ab}$;

2. specifying the functional form of the covariance with distance $k(\tau)$,

These cannot be completely separated; there exists an arbitrary scaling, $\alpha$ of the distance $\tau \to \alpha\tau$ which can be absorbed into the definition of the covariance, $k(\tau) \to k(\tau/\alpha)$. However, provided the steps are tackled in order, there is no ambiguity.

[3] We have yet to define a metric on parameter space with which to take the norm of this vector (see Sec. 11.2.2), but all that is required here is that a suitably smooth metric exists.

### 11.2.1 The metric $g_{ab}$

One simple way to define a distance $\tau$ between two points in parameter space, and the way used in the SE covariance function in Eq. (130), is to define $\tau^2 = g_{ab}(\mathbf{x}_1 - \mathbf{x}_2)^a(\mathbf{x}_1 - \mathbf{x}_2)^b$, where $g_{ab}$ are constant hyperparameters. This distance is obviously invariant under a simultaneous translation of $\mathbf{x}_1 \to \mathbf{x}_1 + \boldsymbol{\Delta}$ and $\mathbf{x}_2 \to \mathbf{x}_2 + \boldsymbol{\Delta}$; therefore, this defines a stationary GP. For a $D$-dimensional parameter space, this involves specifying $D(D+1)/2$ hyperparameters $g_{ab}$.

More complicated distance metrics (with a larger number of hyperparameters) are possible if the condition of stationarity is relaxed, i.e. $g_{ab} \to g_{ab}(\mathbf{x})$. Given a family of stationary covariance functions, a non-stationary generalisation can be constructed. A stationary covariance function can be considered as a kernel function centred at $\mathbf{x}_1$; $k(\mathbf{x}_1, \mathbf{x}_2) \equiv k_{\mathbf{x}_1}(\mathbf{x}_2)$. Allowing a different kernel function to be defined at each point $\mathbf{x}_1$, a new, non-stationary covariance function is $k(\mathbf{x}_1, \mathbf{x}_2) = \int \mathrm{d}\vec{u}\, k_{\vec{u}}(\vec{\lambda_1}) k_{\vec{u}}(\mathbf{x}_2)$.[4] Applying this procedure to a $D$-dimensional SE function generates a non-stationary analogue

$$\begin{aligned} k(\mathbf{x}_i, \mathbf{x}_j) \;&=\; \sigma_f \left|\mathcal{G}^i\right|^{1/4} \left|\mathcal{G}^j\right|^{1/4} \left|\frac{\mathcal{G}^i + \mathcal{G}^j}{2}\right|^{-1/2} \\ &\quad \times \exp\left(-\frac{1}{2} Q_{ij}\right) , \end{aligned} \tag{139}$$

where

$$Q_{ij} \;=\; (\mathbf{x}_i - \mathbf{x}_j)^a (\mathbf{x}_i - \mathbf{x}_j)^b \left(\frac{\mathcal{G}^i_{ab} + \mathcal{G}^j_{ab}}{2}\right)^{-1} , \tag{140}$$

and $\mathcal{G}^i_{ab} = \mathrm{inv}[g_{ab}(\mathbf{x}_i)]$ is the inverse of the parameter-space metric at position $\mathbf{x}_i$. Provided that the metric $g_{ab}(\mathbf{x})$ is smoothly parameterised this non-stationary SE function retains the smoothness properties discussed earlier.

The generalisation in Eq. (139) involves the inclusion of a large set of additional hyperparameters to characterise how the metric changes over parameter space; for example one possible parameterisation would be the Taylor series

$$g_{ab}(\mathbf{x}) = g_{ab}(\mathbf{x}_0) + (\mathbf{x}^c - \mathbf{x}^c_0) \left.\frac{\partial g_{ab}(\mathbf{x})}{\partial \lambda^c}\right|_{\mathbf{x}=\mathbf{x}_0} + \ldots \tag{141}$$

with the hyperparameters $g_{ab}(\mathbf{x}_0)$, $\partial g_{ab}(\mathbf{x})/\partial\lambda^c$, and so on. The inclusion of even a single extra hyperparameter can incur a significant Occam penalty which pushes the training set to favour a simpler choice of covariance function. For this reason most applications use stationary GPs.

An alternative to considering non-stationary metrics is instead to try and find new coordinates $\tilde{\lambda} \equiv \tilde{\lambda}(\mathbf{x})$ such that the metric in these coordinates becomes (approximately) stationary. Such transformations are very problem specific and finding them typically requires expert knowledge of the context of the application.

[4] To see that $k$ is a valid covariance function consider an arbitrary series of points $\{\mathbf{x}_i\}$, and the sum over training set points $I = \sum_{i,j} a_i a_j k(\mathbf{x}_i, \mathbf{x}_j)$; for $k$ to be a valid covariance it is both necessary and sufficient that $I \geq 0$. Using the definition of $k$ gives $I = \int \mathrm{d}\vec{u} \sum_{i,j} a_i a_j k_{\vec{u}}(\vec{\lambda_i}) k_{\vec{u}}(\mathbf{x}_j) = \int \mathrm{d}\vec{u}\, (\sum_i a_i k_{\vec{u}}(\vec{\lambda_i}))^2 \geq 0$.

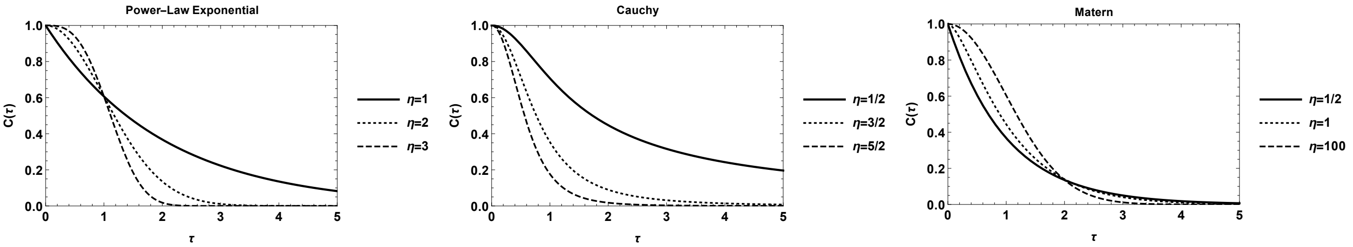

Figure 47: Plots of the different generalisations of the SE covariance function discussed in Sec. 11.2.2. The left-hand panel shows the PLE function, the centre panel shows the Cauchy function, and the right-hand panel shows the Matérn function; in all cases the value of $\sigma_f$ was fixed to unity. In each panel the effect of varying the additional hyperparameter is shown by the three curves. For the PLE covariance the case $\eta = 2$ recovers the SE covariance, while for the Cauchy and Matérn covariances the case $\eta \to \infty$ recovers the SE covariance.

### 11.2.2 The functional form of $k(\tau)$

The second stage of specifying the covariance function involves choosing the function of distance $k(\tau)$. In general whether a particular function $k(\tau)$ is positive definite (and hence is a valid covariance function) depends on the dimensionality $D$ of the underlying space (i.e. $\mathbf{x} \in \mathbb{R}^D$); however, all the functions considered in this section are valid for all $D$. Several choices for $k(\tau)$ are particularly common in the literature, these include the SE covariance function (which has already been introduced), given by

$$k_{\mathrm{SE}}(\tau) = \sigma_f^2 \exp\left(-\frac{1}{2}\tau^2\right) . \tag{142}$$

The power-law exponential (PLE) covariance function, given by

$$k_{\mathrm{PLE}}(\tau) = \sigma_f^2 \exp\left(-\frac{1}{2}\tau^\eta\right) , \tag{143}$$

where $0 < \eta \leq 2$. The PLE reduces to the SE in the case $\eta = 2$. The Cauchy function, given by

$$k_{\mathrm{Cauchy}}(\tau) = \frac{\sigma_f^2}{(1 + \tau^2/2\eta)^\eta} , \tag{144}$$

where $\eta > 0$. This recovers the SE function in the limit $\eta \to \infty$. And finally, the Matérn covariance function, given by

$$k_{\mathrm{Mat}}(\tau) = \frac{\sigma_f^2 2^{1-\eta}}{\Gamma(\eta)} \left(\sqrt{2\eta}\,\tau\right)^\eta K_\eta\left(\sqrt{2\eta}\,\tau\right) , \tag{145}$$

where $\eta > 1/2$, and $K_\eta$ is the modified Bessel function of the second kind [?]. In the limit $\eta \to \infty$, the Matérn covariance function also tends to the SE.

Fig. 47 shows the functional forms of the covariance functions. They have similar shapes; they return a finite covariance at zero distance which decreases monotonically, and tends to zero as the distance becomes large. In the case of regression this indicates that the values of the field at two nearby points in parameter space are closely related, whereas the values at two well separated points are nearly independent. The PLE, Cauchy and Matérn function can all be viewed as attempts to generalise the SE with the inclusion of one extra hyperparameter $\eta$, to allow for more flexible GP modelling. All three alternative functions are able to recover the SE in some limiting case, but the Matérn is the most flexible of the three. This can be seen from the discussion of the MS differentiability of GPs given in section 11.3.

The SE covariance function is infinitely differentiable at $\tau = 0$, and so the corresponding GP is infinitely MS differentiable. The PLE function is infinitely differentiable at $\tau = 0$ for the SE case when $\eta = 2$, but for all other cases it is not at all MS differentiable. In contrast, the Cauchy function is infinitely differentiable at $\tau = 0$ for all choices of the hyperparameter $\eta$. The Matérn function, by contrast, has a variable level of differentiability at $\tau = 0$, controlled via the hyperparameter $\eta$. The GP corresponding to the Matérn covariance function in Eq. (145) is $\zeta$-times MS differentiable if and only if $\eta > \zeta$. This ability to modify the differentiability allows the same covariance function to successfully model a wide variety of data. In the process of maximising the hyperlikelihood for the training set over hyperparameter $\eta$, the GP learns the (non)smoothness properties favoured by the data, and the the GPR returns a correspondingly (non)smooth function.

### 11.2.3 Compact support and sparseness

All of the covariance functions considered up until this point have been strictly positive;

$$k(\tau) > 0 \quad \forall \tau \in [0, \infty)\,. \tag{146}$$

When evaluating the covariance matrix for the training set $K_{ij}$ this leads to a matrix where all entries are positive definite; i.e. a dense matrix. When performing the GPR it is necessary to maximise the hyperlikelihood for the training set with respect to the hyperparameters. This process involves inverting the dense matrix $K_{ij}$ at each iteration of the optimisation algorithm. Although this procedure is carried out offline, it still can become prohibitive for large training sets. A related problem, as pointed out in Sec. **??** is that for large training sets the determinant of the covariance matrix is typically small which also contributes to making the covariance matrix hard to invert.

One potential way around these issues is to consider a covariance function with compact support,

$$\begin{aligned} k(\tau) > 0 \quad & \tau \in [0, T]\,, \\ k(\tau) = 0 \quad & \forall \tau \in (T, \infty)\,, \end{aligned} \tag{147}$$

where $T$ is some threshold distance beyond which we assume that the waveform differences become uncorrelated. This leads to a sparse, band-diagonal covariance matrix, which is much easier to invert. Care must be taken when specifying the covariance function to ensure that the function is positive definite (which is required of a GP): if the SE covariance function is truncated, then the matrix formed from the new covariance function is not guaranteed to be positive definite.

Nevertheless, it is possible to construct covariance functions which have the requisite properties and satisfy the compact support condition in Eq. (147). These are typically based on polynomials. We consider a series of polynomials, originally proposed by Wendland. These have the property that they are positive definite in $\mathbb{R}^D$ and are $2q$-time differentiable at the origin. Therefore the discrete parameter $q$ is in some sense analogous to the $\eta$ hyperparameter of the Matérn covariance function in that it controls the smoothness of the GP.

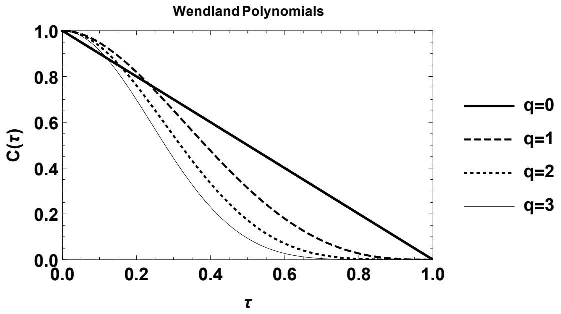

Figure 48: Plots of the first few Wendland polynomial covariance functions. All these functions have compact support, $k(\tau) = 0$ for $\tau > 1$. As the value of $q$ increases the functions become smoother near the origin.

Defining $\beta$ to be

$$\beta = \left\lfloor \frac{D}{2} \right\rfloor + q + 1 \tag{148}$$

and where $\Theta(x)$ denotes the Heaviside step function, the first few Wendland polynomials $k_{D,q}(\tau)$ are given by,

$$\begin{aligned}
k_{D,0}(\tau) &= \sigma_f^2 \Theta(1-\tau)(1-\tau)^{\beta}\,, && (149)\\
k_{D,1}(\tau) &= \sigma_f^2 \Theta(1-\tau)(1-\tau)^{\beta+1}\left[1 + (\beta+1)\,\tau\right]\,, && (150)\\
k_{D,2}(\tau) &= \frac{\sigma_f^2}{3}\Theta(1-\tau)(1-\tau)^{\beta+2}\left[\,3 + (3\beta+6)\,\tau\right.\\
&\quad \left. + \left(\beta^2 + 4\beta + 3\right)\tau^2\right]\,, && (151)\\
k_{D,3}(\tau) &= \frac{\sigma_f^2}{15}\Theta(1-\tau)(1-\tau)^{\beta+3}\left[15 + (15\beta+45)\,\tau\right.\\
&\quad + \left(6\beta^2 + 36\beta + 45\right)\tau^2\\
&\quad \left. + \left(\beta^3 + 9\beta^2 + 23\beta + 15\right)\tau^3\right]\,. && (152)
\end{aligned}$$

The first few Wendland polynomials are plotted in Fig. 48. Other types of covariance functions with compact support have also been proposed and explored in the literature, but we do not consider them here.

## 11.3 Continuity and differentiability of GPs

Before moving on to some examples, we give proofs concerning the continuity and differentiability of GPs. Let $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3 \ldots$ be a sequence of points in parameter space which converges to a point $\mathbf{x}_*$, in the sense $\lim_{\ell\to\infty} \, |\mathbf{x}_\ell \, - \, \mathbf{x}_*| = 0$. The GP $Y(\mathbf{x})$ is said to be MS continuous at $\mathbf{x}_*$ if

$$\lim_{\ell\to\infty} \mathbb{E}\left[(Y(\mathbf{x}_\ell) - Y(\mathbf{x}_*)|Y(\mathbf{x}_\ell) - Y(\mathbf{x}_*))\right] = 0\,, \tag{153}$$

where $\mathbb{E}[\ldots]$ denotes the expectation of the enclosed quantity over realisations of the GP. MS continuity implies continuity in the mean,

$$\lim_{\ell\to\infty} \mathbb{E}\left[Y(\mathbf{x}_\ell) - Y(\mathbf{x}_*)\right] = 0\,. \tag{154}$$

This follows from considering the variance of the quantity $Y(\mathbf{x}_\ell) - Y(\mathbf{x}_*)$, and the fact that variance is non-negative. There are other notions of continuity of GPs used in the literature, but the notion of MS continuity relates most easily to the covariance.

The mean and the covariance of a GP are defined as

$$\begin{aligned} m(\mathbf{x}) &= \mathbb{E}[Y(\mathbf{x})]\,, \\ k(\mathbf{x}_1, \mathbf{x}_2) &= \mathbb{E}[(Y(\mathbf{x}_1) - m(\mathbf{x}_1)|Y(\mathbf{x}_2) - m(\mathbf{x}_2))]\,. \end{aligned} \tag{155}$$

Using these, Eq. (153) can be written as

$$\begin{aligned} \lim_{\ell\to\infty} \{k(\mathbf{x}_*, \mathbf{x}_*) - 2k(\mathbf{x}_\ell, \mathbf{x}_*) + k(\mathbf{x}_\ell, \mathbf{x}_\ell) & \\ + (m(\mathbf{x}_*) - m(\mathbf{x}_\ell)|m(\mathbf{x}_*) - m(\mathbf{x}_\ell))\} &= 0\,, \end{aligned} \tag{156}$$

and using the continuity of the mean in Eq. (154) gives

$$\lim_{\ell\to\infty} [k(\mathbf{x}_*, \mathbf{x}_*) - 2k(\mathbf{x}_\ell, \mathbf{x}_*) + k(\mathbf{x}_\ell, \mathbf{x}_\ell)] = 0\,. \tag{157}$$

This condition is satisfied if the covariance function is continuous at the point $\mathbf{x}_1 = \mathbf{x}_2 = \mathbf{x}_*$. Therefore, we arrive at the result that if the covariance function is continuous in the usual sense at some point $\mathbf{x}_*$, then the corresponding GP is MS continuous at this point.[5] In the special case of stationary covariance this reduces to checking continuity of $k(\vec{\tau})$ at $\vec{\tau} = 0$, and in the special case of isotropic covariance, continuity of $k(\tau)$ at $\tau = 0$.

We now move on from continuity to consider differentiability. In the spirit of Eq. (153), the notion of taking the MS derivative of a GP is defined as

$$\frac{\partial Y(\mathbf{x})}{\partial \mathbf{x}^a} = \underset{\epsilon\to 0}{\text{l.i.m}}\, X_a(\mathbf{x}, \epsilon)\,, \tag{158}$$

where l.i.m is read limit in MS and

$$X_a(\mathbf{x}, \epsilon) = \frac{Y(\mathbf{x} + \epsilon\,\hat{e}_a) - Y(\mathbf{x})}{\epsilon} \tag{159}$$

with parameter-space unit vector $\hat{e}_a$. This definition can be extended to higher-order derivatives in the obvious way.

The MS derivative of a GP is also a GP; this follows simply from the fact that the sum of Gaussians is also distributed as a Gaussian. The covariance of $X_a(\mathbf{x}, \epsilon)$ is given by

$$\begin{aligned} K_\epsilon(\mathbf{x}_1, \mathbf{x}_2) &= \mathbb{E}\left[(X_a(\mathbf{x}_1, \epsilon) - \Xi(\mathbf{x}_1, \epsilon)|\right. \\ &\qquad \left. X_a(\mathbf{x}_2, \epsilon) - \Xi(\mathbf{x}_2, \epsilon))\right] \end{aligned} \tag{160}$$

where $\Xi_a(\mathbf{x}, \epsilon) = \mathbb{E}[X_a(\mathbf{x}, \epsilon)]$. It then follows that

$$\begin{aligned} K_\epsilon(\mathbf{x}_1, \mathbf{x}_2) &= \frac{k(\mathbf{x}_1 + \epsilon, \mathbf{x}_2 + \epsilon) - k(\mathbf{x}_1, \mathbf{x}_2 + \epsilon)}{\epsilon^2} \\ &\quad + \frac{k(\mathbf{x}_1 + \epsilon, \mathbf{x}_2) - k(\mathbf{x}_1, \mathbf{x}_2)}{\epsilon^2}\,. \end{aligned} \tag{161}$$

[5] A GP is continuous in MS if <u>and only if</u> the covariance function is continuous, although this is not proved here.

Substituting this into Eq. (158), the limit in MS becomes a normal limit, and the result is obtained that the MS derivative of a MS continuous GP with covariance function $k(\mathbf{x}_1, \mathbf{x}_2)$ is a GP with covariance function $\partial^2 k(\mathbf{x}_1, \mathbf{x}_2)/\partial \mathbf{x}_1^a \partial \mathbf{x}_2^a$. In general the covariance function of the $\zeta$-times MS differentiated GP

$$\frac{\partial^\zeta Y(\mathbf{x})}{\partial \mathbf{x}^{a_1} \partial \mathbf{x}^{a_2} \dots \partial \mathbf{x}^{a_\zeta}}\,, \tag{162}$$

is given by the $2\zeta$-times differentiated covariance function

$$\frac{\partial^{2\zeta} k(\mathbf{x}_1, \mathbf{x}_2)}{\partial \mathbf{x}_1^{a_1} \partial \mathbf{x}_2^{a_1} \partial \mathbf{x}_1^{a_2} \partial \mathbf{x}_2^{a_2} \dots \partial \mathbf{x}_1^{a_\zeta} \partial \mathbf{x}_2^{a_\zeta}}\,. \tag{163}$$

From the above results relating the MS continuity of GPs to the continuity of the covariance function at $\mathbf{x}_1 = \mathbf{x}_2 = \mathbf{x}_*$, it follows that the $\zeta$-times MS derivative of the GP is MS continuous (the GP is said to be $\zeta$-times MS differentiable) if the $2\zeta$-times derivative of the covariance function is continuous at $\mathbf{x}_1 = \mathbf{x}_2 = \mathbf{x}_*$. So it is the smoothness properties of the covariance function along the diagonal points that determine the differentiability of the GP.[6]

## 11.4 Example applications of Gaussian processes

**Example: interpolation of a quadratic** We consider first a toy problem in which we generate noisy measurements, $\{y_i\}$, at 200 points, $\{x_i\}$, randomly chosen in the interval $[0, 1]$ according to

$$y_i = -2 - 3x_i + 5x_i^2 + \epsilon_i, \qquad \epsilon_i \sim N(0, 0.15^2).$$

We then fit a Gaussian process to a training set comprising a subset of these points. We use a squared exponential covariance function and optimize the hyperparameters over the training set. The results of this procedure are shown in Figure 49. Results are represented by the expectation value and $1\sigma$ uncertainty computed from the fitted Gaussian process as a function of $x$. We see that the Gaussian process is well able to recover the true function, even with as a few as ten training points. This is a particularly simple function and if we knew that the relationship was quadratic there would be no need to use a Gaussian process to fit the data. In Figure 50 we show the result of fitting a quadratic model to the same data. As expected, the fit is slightly better, but not hugely so. The advantage of the Gaussian process approach is that you do not need to know the form of the model in advance, and avoid the problem of model mis-specification. In Figure **??** we show the result of fitting a linear model to the same data. We see that we end up with a very precise, but wrong, representation of the curve. Gaussian process regression models have greater flexibility and should always converge to the true underlying function in the limit that the number of observations tends to infinity.

**Example: waveform model errors** We will now consider a few examples from the gravitational wave literature. There are many of these that have all appeared since $\sim$ 2015, so we cannot describe them all but we will mention a few different examples. The first application of Gaussian processes in a gravitational wave context was to characterise uncertainties coming from waveform model errors (Moore & Gair (2014)). A Gaussian process was used to model the error in a particular waveform model family over parameter

[6] It can be further shown that if a covariance function $k(\mathbf{x}_1, \mathbf{x}_2)$ is continuous at every diagonal point $\mathbf{x}_1 = \mathbf{x}_2$ then it is everywhere continuous.

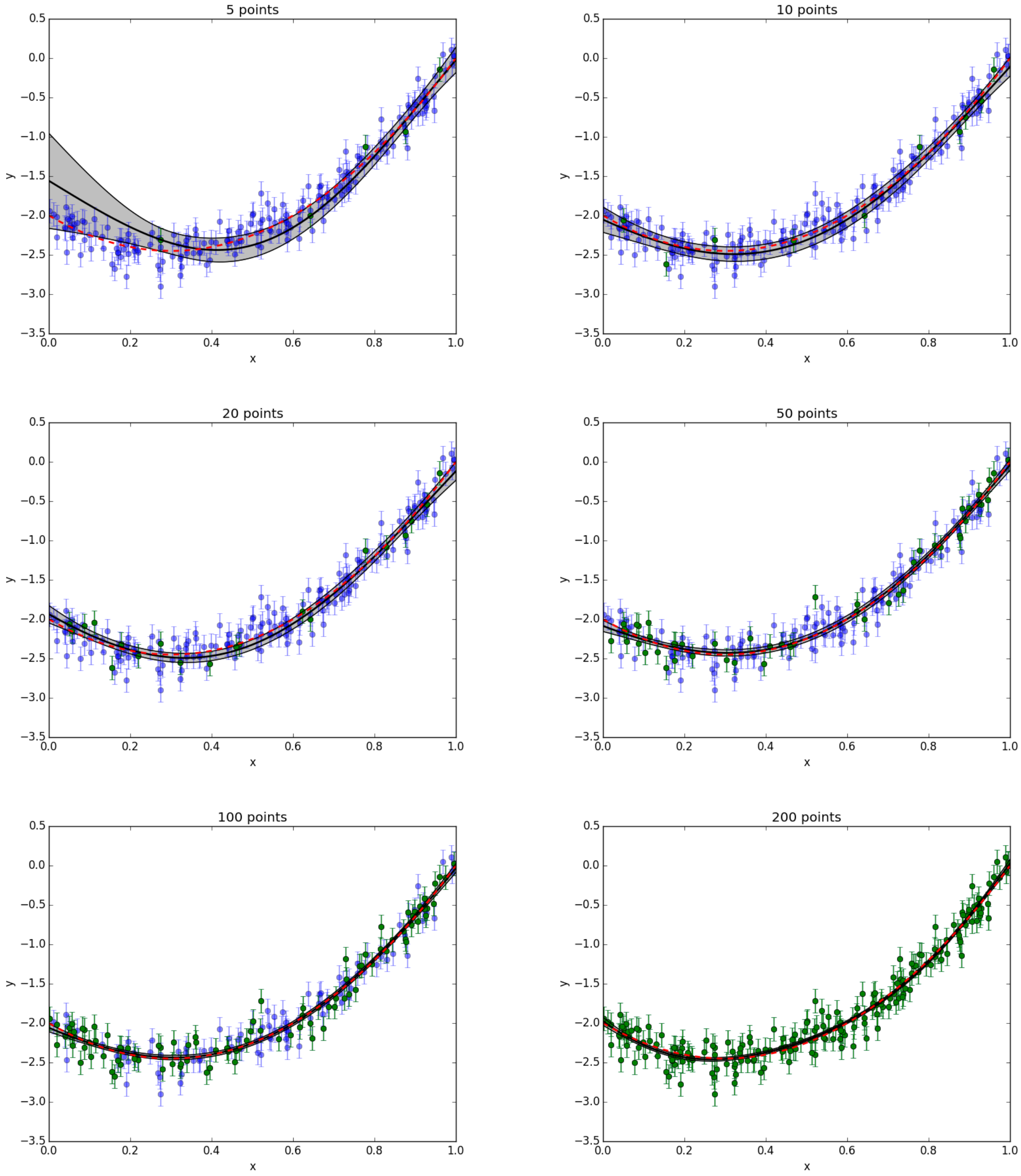

Figure 49: Gaussian process fit to noisy measurements of a quadratic, for different sizes of training set, as stated in the title of each panel.

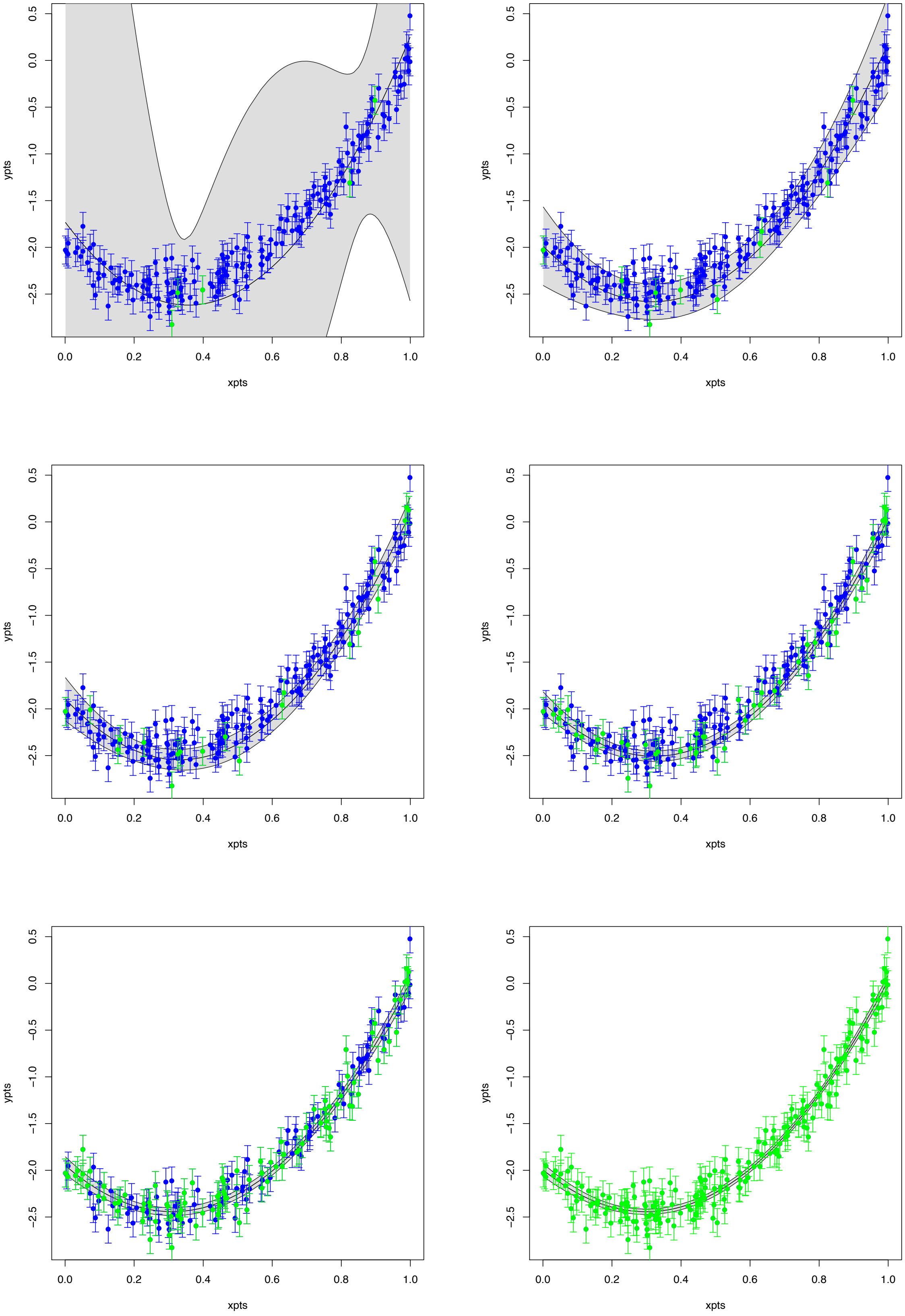

Figure 50: As Figure 49, but now fitting a quadratic linear model to the same data.

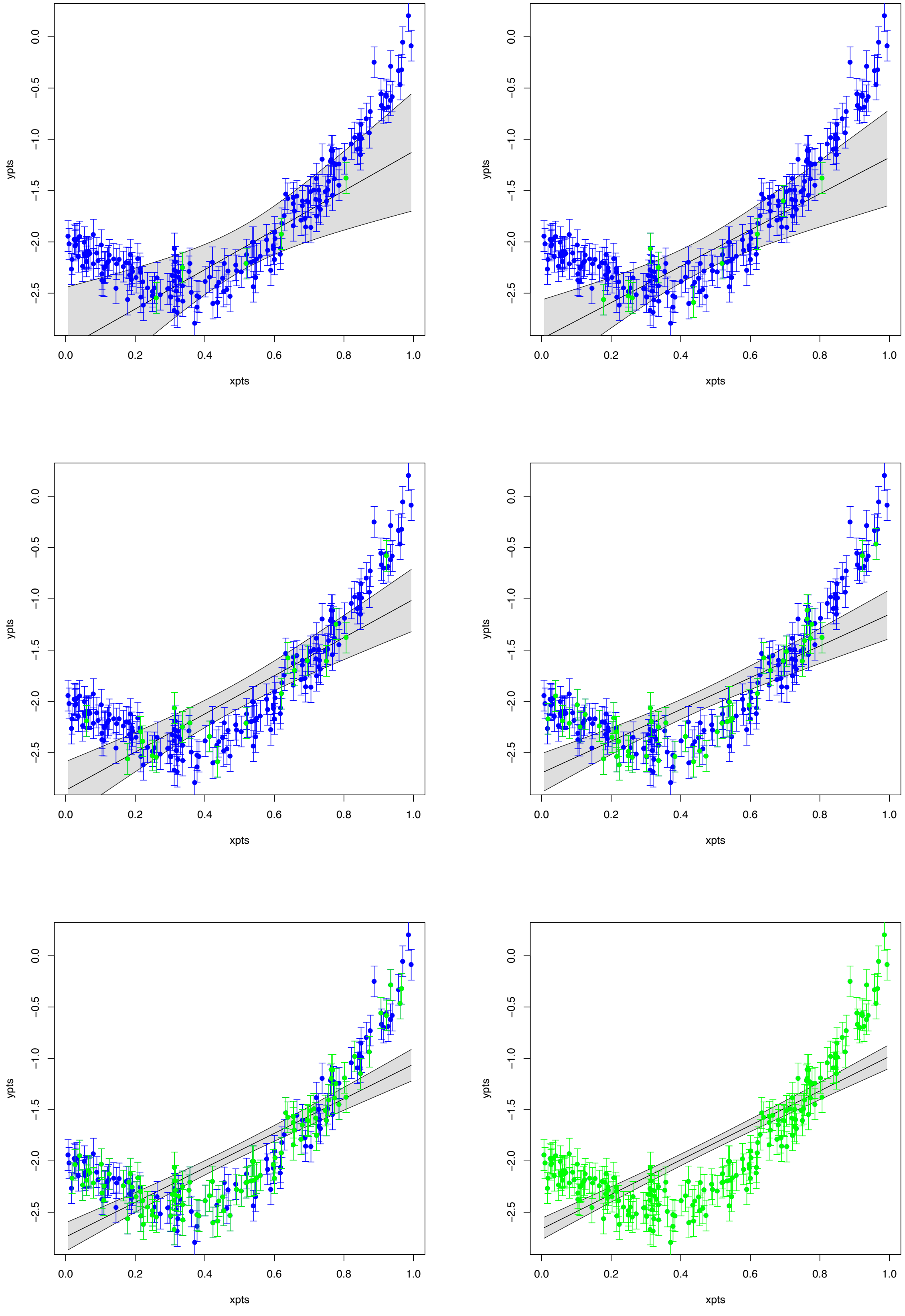

Figure 51: As Figure 49, but now fitting a linear model to the same data.

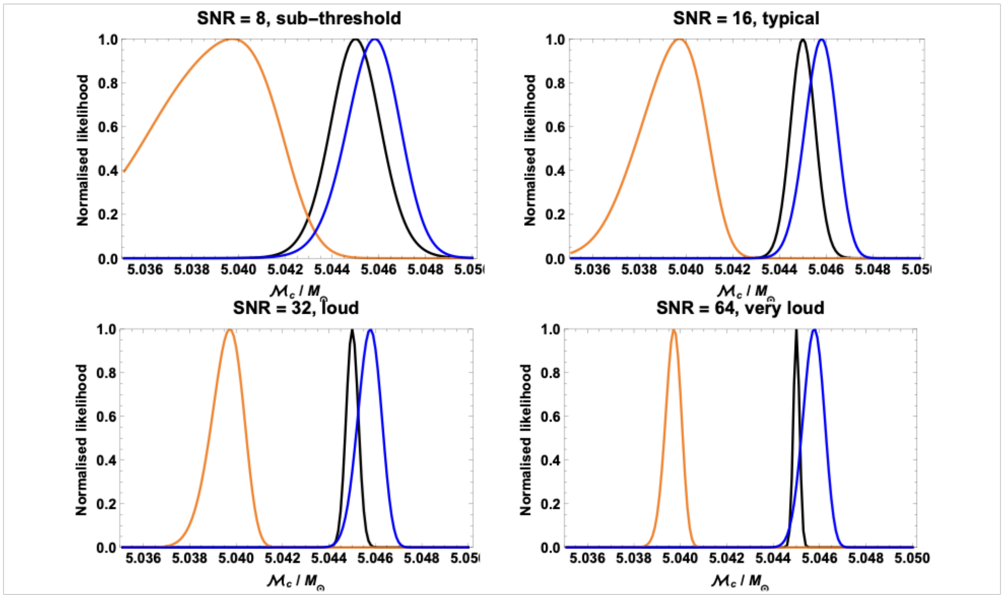

Figure 52: Comparison between uncorrected, corrected and "true" likelihood for inference with waveform models that include model error. The corrected likelihood uses a Gaussian process to model the waveform error and then marginalises this out of the likelihood. Reproduced from Moore et al. (2015).

space. Using a training set based on model errors estimated as the difference between two different approximate waveforms, a Gaussian process model for the waveform error was produced. As this distribution is Gaussian and so is the normal gravitational wave likelihood, the waveform error can then be marginalised out of the likelihood to give an alternative **marginalised likelihood** for use in parameter estimation. This marginalised likelihood took the form

$$\mathcal{L}(\vec{\lambda}) \quad \propto \quad \frac{1}{\sqrt{1+\sigma^2(\vec{\lambda})}} \exp\left( -\frac{1}{2} \frac{\left\| s - H(\vec{\lambda}) + \mu(\vec{\lambda}) \right\|^2}{1+\sigma^2(\vec{\lambda})} \right). \tag{164}$$

In this $\vec{\lambda}$ is the vector of parameters characterising the gravitational wave signal, the quantity $\mu(\vec{\lambda})$ is the Gaussian process estimate for the model error, and shifts the distribution to eliminate the error, and $\sigma^2(\vec{\lambda})$ is the variance in the Gaussian process, which widens the posterior to account for the uncertainty in the model error. Use of this marginalised likelihood corrects for biases in parameter estimation, as illustrated in Figure 52.

**Example: waveform interpolation** In Williams et al. (2020), Gaussian processes were used to directly model the gravitational waveform, rather than its error. A set of numerical relativity waveforms were used to create a training set to which a Gaussian process model was fitted. In Figure 53 we show some random draws from the GP model at a certain point in parameter space and compare these to two different waveform approximants evaluated at the same point. We see that the GP uncertainty band includes all of the different approximants and so automatically factors in waveform uncertainty.

**Example: population inference** In Taylor & Gerosa (2018), a Gaussian process was used as a means to interpolate the output of binary population synthesis code over the

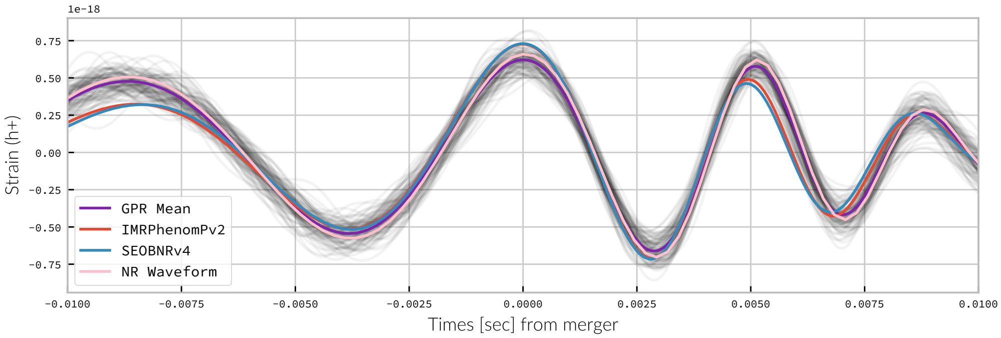

Figure 53: Comparison of several approximate waveform models to random draws from a Gaussian process interpolant trained on numerical relativity simulations. Reproduced from Williams et al. (2020).

space of physical parameters that characterise them. The resulting model, continuous over parameter space, was then used to infer properties of the underlying astrophysical population based on a set of observed compact binary inspirals. Figure **??** shows simulated inferred posteriors on the population parameters that were produced in this way.

**Example: equation of state uncertainties** Landry & Essick (2019) and Essick, Landry & Holz (2019) used a Gaussian process to model the equation of state of a neutron star, $p(\rho)$. The hyperparameters of the Gaussian process were constrained using a training set including numerical equation of state simulations. The resulting model generates random equations of state which can be used to marginalise equation of state uncertainties out of inference on gravitational wave signals from binary neutron stars. Figure 55 shows a set of random draws of the equation of state from the Gaussian process.

## 11.5 Dirichlet Processes

Recall that a Dirichlet distribution generates a set of $K$ random values, $\{x_i\}$, constrained to take values with $0 \leq x_i \leq 1$ for all $i$ and $\sum x_i = 1$. The distribution depends on a vector of parameters $\vec{\alpha} = (\alpha_1, \dots, \alpha_K)$ and has pdf

$$p(\vec{x}) = \frac{1}{B(\vec{\alpha})} \prod_{i=1}^{K} x_i^{\alpha_i - 1}, \qquad B(\vec{\alpha}) = \frac{\prod_{i=1}^{K} \Gamma(\alpha_i)}{\Gamma\left(\sum_{j=1}^{K} \alpha_j\right)}.$$

A realisation of a Dirichlet distribution is a probability mass function for a discrete distribution with $K$ possible outcomes. A **Dirichlet process** generalises the Dirichlet distribution to infinite dimensions and a realisation of a Dirichlet process is a continuous probability distribution. A Dirichlet process is characterised by a **base distribution**, $P$, and a **concentration parameter**, $a$. The base distribution is a probability measure on a set $S$. The process $X$ is a Dirichlet process, denoted $X \sim$DP$(P, a)$ if for any measurable finite partition of the set $S$, $\{B_i\}_{i=1}^{n}$, the probability distribution on this partition generated by $X$ is

$$(X(B_1), X(B_2), \dots, X(B_n)) \sim \text{Dir}(aP(B_1), aP(B_2), \dots, aP(B_n)). \qquad (165)$$

In the limit $a \to 0$, the Dirichlet pdf, which is proportional to $x_i^{\alpha_i - 1}$, places a logarithmically increasing weight towards the lower boundary of the variable range. Draws from this

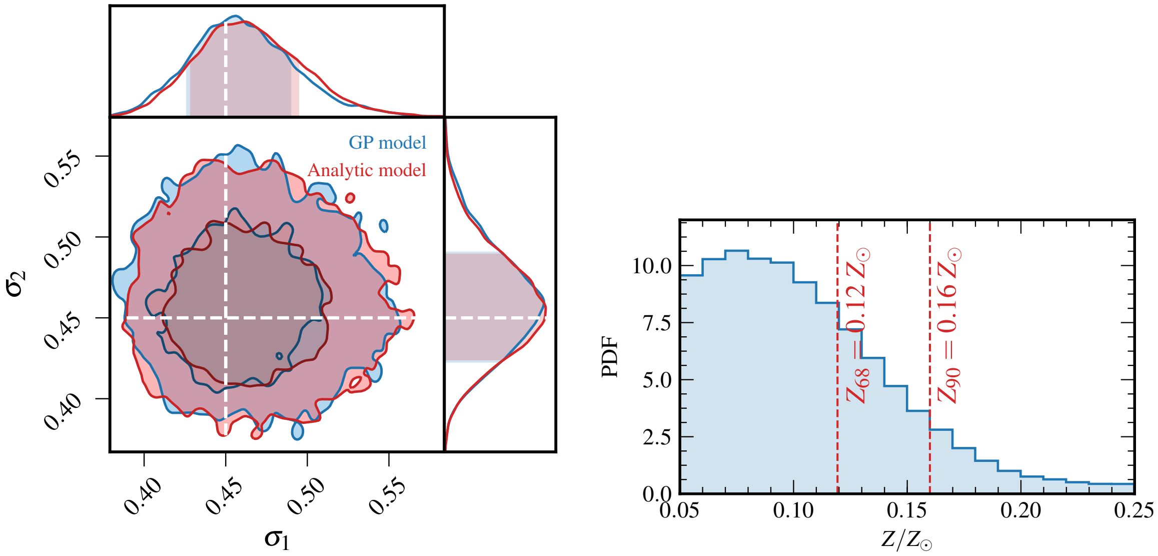

Figure 54: Posteriors on physical parameters of the astrophysical source population inferred form simulated observations of binaries. Inference relied on a Gaussian process model that interpolated the output of the population synthesis codes over the astrophysical parameter space. Reproduced from Taylor & Gerosa (2018).

distribution will therefore be singletons, with all $x_i$'s bar one equal to zero. For small $a$ the Dirichlet distribution will therefore tend to be discretized, with probability concentrated at a small number of locations.

In the limit $a \to \infty$, the distribution becomes more and more concentrated at its mode, which is at $x_i = P(B_i)$. Every realisation of $\text{Dir}(aP(B_1), aP(B_2), \ldots, aP(B_n))$ therefore returns $(P(B_1), \ldots, P(B_n))$ and every realisation of the Dirichlet process thus gives the base distribution.

These limits show that the Dirichlet process generates discretized representations of the base distribution, with the level of discretization decreasing as $a \to \infty$. To illustrate this, we show in Figure 56 and 57 some realisations of a Dirichlet process, for a fixed base distribution, $P = N(0, 1)$, and various choices of $a$. In each figure, we represent the realisation of the Dirichlet process by a set of 1000 random draws from the realised probability distribution. It is clear that for small $a$, only a small number of values are returned, showing high discretisation, but as $a$ increases the number of distinct values is increasing and the distribution becomes a closer and closer approximation to the base distribution.

### 11.5.1 Sampling Dirichlet processes

A realisation of a Dirichlet process is a probability distribution on $S$ and hence infinite dimensional. Drawing such a realisation is therefore very difficult. However, in practice what we need is not the realisation of the Dirichlet process itself but a set of samples from that realised distribution, which is much easier to obtain. If the full realisation is required, this can be evaluated by looking at the distribution of a large number of samples. This is how the realisations shown in Figures 56 and 57 were produced.

There are several different algorithms for drawing samples from a random realisation of a Dirichlet process, $X \sim$DP$(P, a)$. The **chinese restaurant process** generates a sequence

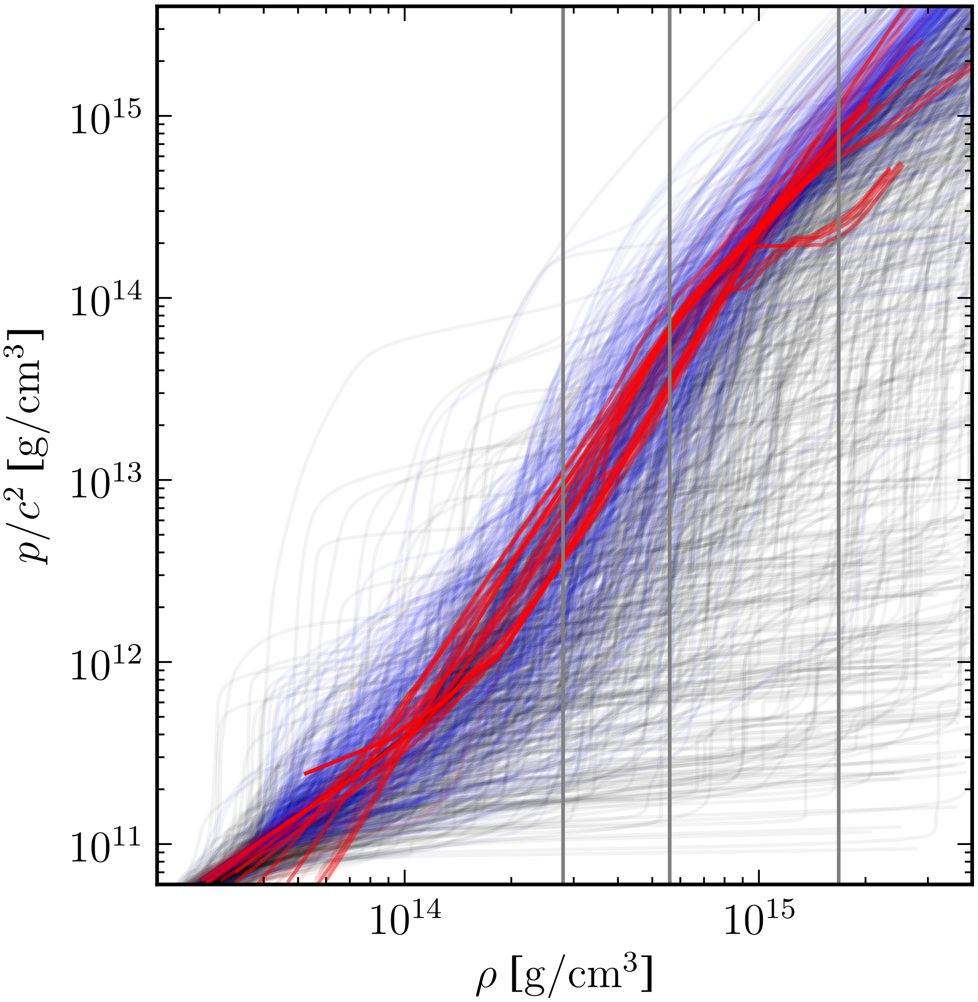

Figure 55: Random draws from a Gaussian process model of the equation of state of a neutron star. Reproduced from Essick et al. (2019).

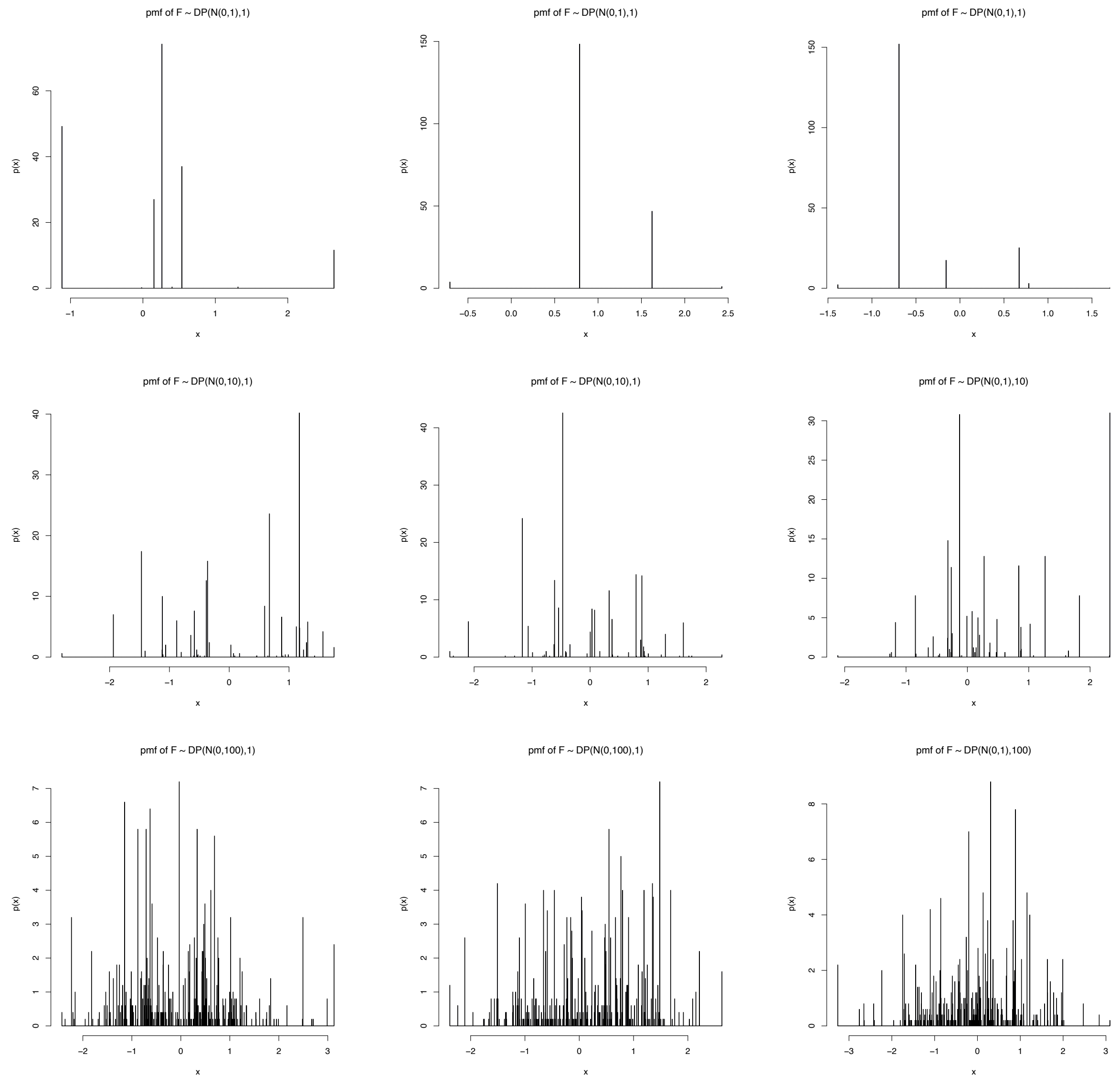

Figure 56: Sample realisations of a Dirichlet process, $X \sim \mathrm{DP}(N(0,1),a)$, for $a = 1$ (top row), $a = 10$ (middle row) and $a = 100$ (bottom row). In each figure we show 1000 samples from the given realisation of the Dirichlet process. Within each row, the figures show three distinct realisations of the stated Dirichlet process.

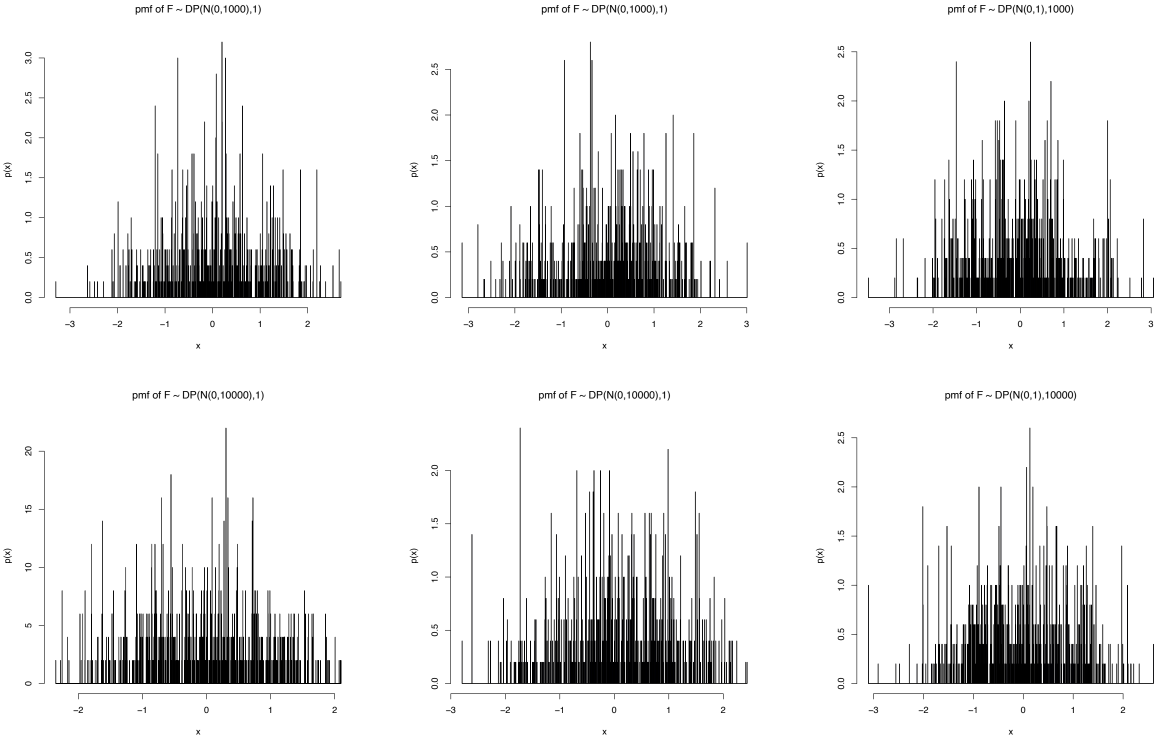

Figure 57: As in Figure 56, these figures show sample realisations of a Dirichlet process, $X \sim$DP$(N(0,1),a)$, for $a = 1000$ (top row) and $a = 10000$ (bottom row). In each figure we show 1000 samples from the given realisation of the Dirichlet process. Within each row, the figures show three distinct realisations of the stated Dirichlet process.

of samples $\{x_i\}$ for $i \geq 1$ as follows

- with probability $a/(a+i-1)$ draw $x_i$ from P;
- with probability $n_x/(a+i-1)$ set $x_i = x$, where $n_x$ is the number of previous observations of $x_j = x$ for $j < i$.

This procedure is called the chinese restaurant process by analogy with a restaurant with an infinite number of tables, each serving a different dish, and each with infinite seating capacity. A new diner may choose to sit at a new table, or may choose to sit at a table where people are already eating. The probability of choosing a particular table is proportional to the number of people observed already sitting at that table and enjoying the offered dish.

Closely related to this is the **Polya Urn** scheme. In that construction we start with an urn containing $a$ black balls. At each step of the algorithm, a ball is drawn at random from the urn. If the ball is black, we generate a new color randomly, color a new ball this color and return it to the urn along with the black ball. The corresponding sample is the new color. If the ball drawn is coloured, then we take a new ball, color it the same color as the sampled ball, and return both of them to the urn. The corresponding sample is the color of the ball that was drawn. It is clear that the distribution of colors produced in this way corresponds to the samples generated form the chinese restaurant process.

A final approach to constructing a sample from a random realisation of a Dirichlet process is the **stick breaking** construction. This approach explicitly generates a discrete distribution, $X$, which is a realisation of the Dirichlet process. The distribution is given by

$$\begin{aligned}
&X = \left(\sum_{l=1}^{L_H} p_l \delta_{U_l}\right) + \left(1 - \sum_{l=1}^{L_H} p_l\right) \delta_{U_0} \\
&p_1 = V_1, \qquad p_l = \left(\prod_{j=1}^{l-1} (1 - V_j)\right) V_l, \quad l \geq 2, \qquad p_0 = 1 - \sum_{l=1}^{L_H} p_l \\
&V_l \sim \text{Beta}(1, a), \quad l = 1, \ldots, L_H, \qquad U_l \sim P, \quad l = 0, 1, \ldots, L_H,
\end{aligned} \tag{166}$$

where we take the limit $L_H \to \infty$, but in practical applications the procedure is truncated at some finite, but sufficiently large, value.

### 11.5.2 Example applications

The main application of Dirichlet processes is in the field of Bayesian nonparametrics, where they are used as a prior for unknown probability distributions. We will provide two examples.

**Example: B-spline regression** In the nonparametric regression chapter we encountered the notion of smoothing splines for regression. In that context, the knots of the spline were fixed at the locations of the observed data points. The number of knots is therefore fixed for any given data set and grows as $n \to \infty$. The smoothing was controlled by the regularisation parameter. Another approach to nonparametric regression is to allow the number of spline points to vary and let the data choose the optimal number. Even greater flexibility comes from allowing the locations of the spline knots to vary. In Edwards & Gair (2020) they presented a Bayesian nonparametric regression algorithm that uses B-splines (an alternative basis for cubic splines than the one presented in this course), but with the number and location of the knots both allowed to vary and adapt to the data. The knot locations were

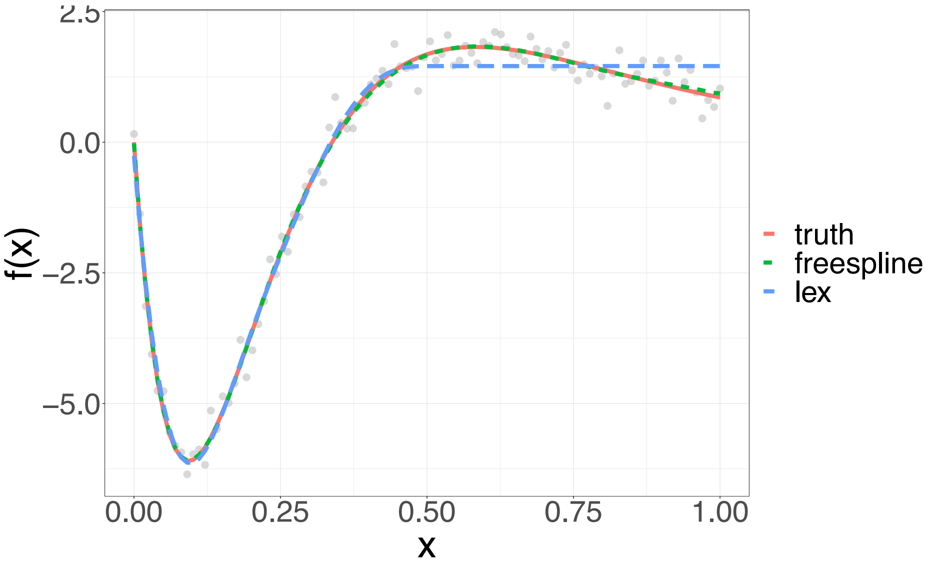

Figure 58: Nonparametric regression fit to noisy measurements of the function $f(x) = 26\exp(-3.25x) - 4\exp(-6.5x) + 3\exp(-9.75x)$ using the freespline algorithm with a Dirichlet process prior on the probability density determining the knot locations. Figure reproduced from Edwards & Gair (2020).

represented by a random cumulative density function, $H$, defined on the interval $[0, 1]$, with the $j$'th of $k - r$ internal knots located at $x_j = H(j/(k - r))$. The random density $H$ was assigned a Dirichlet process prior. In Figure 58 we show the result of using this algorithm to fit noisy measurements of a function

$$f(x) = 26\exp(-3.25x) - 4\exp(-6.5x) + 3\exp(-9.75x).$$

We see that the freespline algorithm is able to capture all of the turning points of this function, while another widely used regression algorithm, LEX, is not. In Figure 59 we show another application of that algorithm to obtain a nonparametric fit to the power spectrum of temperature fluctuations in the CMB measured by Planck. The nonparametric fit can be compared to the best fit cosmological model prediction. There is some evidence that the data does not support the up-tick at low multipoles predicted by the model. In fact, there has been extensive debate in the literature about whether the $l = 2, 3$ multipoles are in fact lower than predicted, and these results seem to support that. There is also weak evidence that the data suggests the second and third peaks are further apart than the standard ΛCDM model predicts. Observations of this nature (if they were to be robust in future data sets) would help guide modifications to the model, and this would be much harder without the nonparametric regression tool.

**Example: LIGO sky localisation** In Del Pozzo et al. (2018), a Dirichlet process Gaussian mixture model (DPGMM) was used to produce a smooth interpolation of the output of LALInference sampling. The aim was to produce a continuous representation of the source localisation volume (sky location and distance), to target electromagnetic follow-up. The Dirichlet process was used as a prior to generate the centres (in 3-dimensions) of Gaussians. The sum of these Gaussians, with weights, was used as a representation of the smooth posterior probability and then constrained by the set of posterior samples

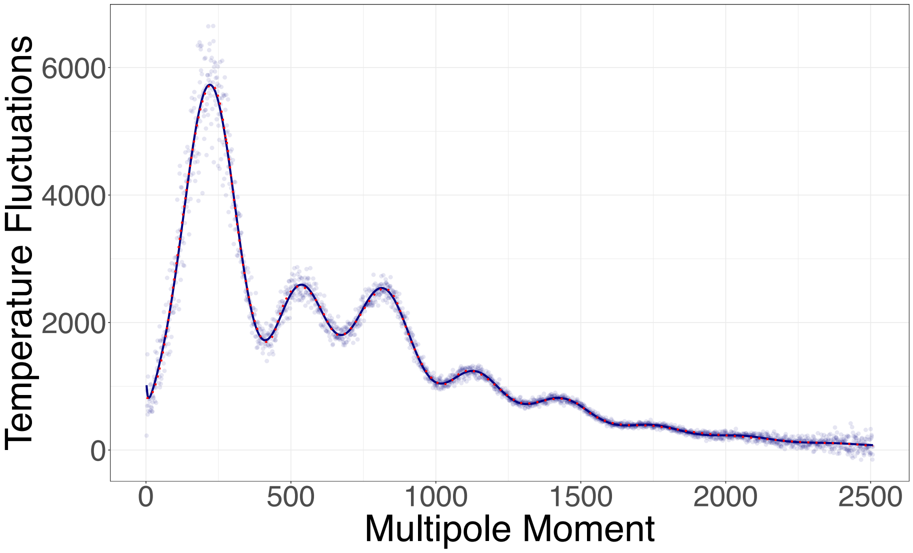

Figure 59: Nonparametric regression fit to the CMB temperature power spectrum, as measured by Planck. The dashed red line is the freespline fit to the data, while the blue line is the prediction of the best fit cosmological model. Figure reproduced from Edwards & Gair (2020).

previously generated by LALInference. In Figure 60 we show the result of this analysis, the distribution of posterior credible volumes computed for a set of injections and using the DPGMM to obtain the credible volumes. This is the only application of Dirichlet processes in a gravitational wave context to date, but they are likely to be powerful tools for fitting nonparametric population models as the number of observations becomes large enough to make this possible.

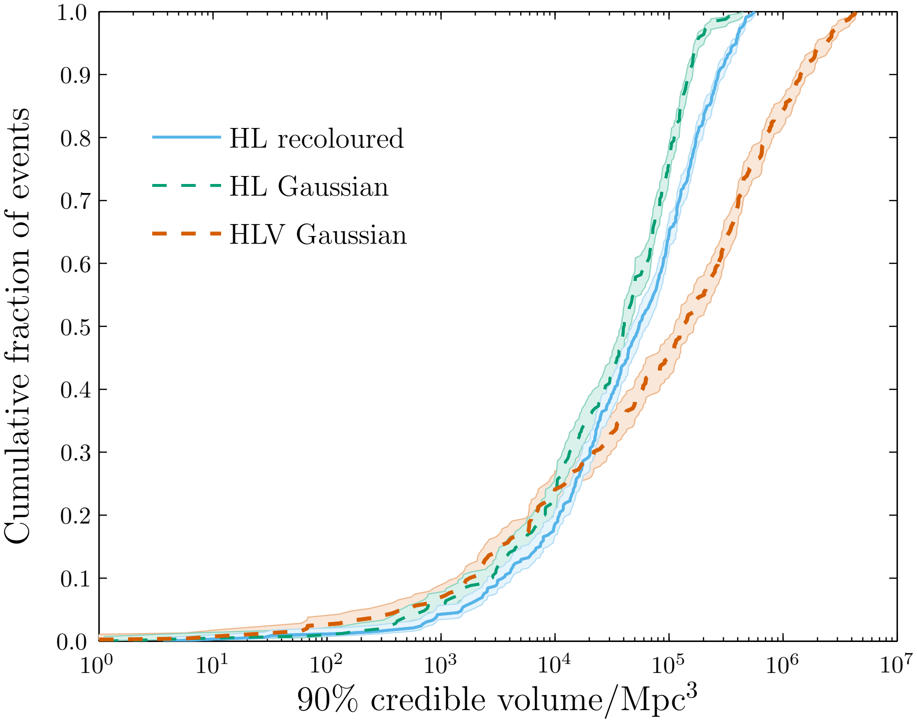

Figure 60: Cumulative distribution function of 90% credible volumes for events observed by the ground-based detector network. The credible volumes were computed by fitting a Dirichlet Process Gaussian Mixture Model to posterior samples generated by LALInference. Figure reproduced from Del Pozzo et al. (2018).