Conditioning on Outputs of Linear Operators

Suppose we have a function $f\colon \mathcal{X} \to \mathbb{R}$ with a Gaussian process prior distribution:

$$p(f) = \mathcal{GP}(f; \mu, K).$$

We have discussed how to perform inference about f when given (noisy) observations of the function at a set of points $\mathbf{X}$: $\mathcal{D} = (\mathbf{X}, \mathbf{y})$. Here we are going to expand the types of observations we may use during GP inference.

Functionals and linear functionals

Specifically, we are going to consider so-called *linear functionals* of f. A *functional* is a function $L[f]$ that takes as an input a function f and returns a scalar. (Functionals are sometimes called "functions of functions.") A very simple example of a functional is the *point-evaluation functional.* Let $x \in \mathcal{X}$ be an arbitrary fixed point in the domain. We define a corresponding functional L_x by

$$f \mapsto L_x[f] = f(x).$$

So, given a function f, the point-evaluation functional L_x simply evaluates f at x and returns the result. This is a functional we are very accustomed to using.

A functional is said to be *linear* when it satisfies a simple linearity property. Specifically, let $a \in \mathbb{R}$ be an arbitrary scalar constant and let f and g be two arbitrary functions. A functional L is linear if the following equality always holds:

$$L[af + g] = aL[f] + L[g].$$

It is easy to see that the point-evaluation functional L_x is linear:

$$L_x[af + g] = (af + g)(x) = af(x) + g(x) = aL_x[f] + L_x[g].$$

There are several other quite-common linear functionals that we are familiar with. The two we will discuss here are integration against an arbitrary function $p(x)$:

$$f \mapsto I_p[f] = \int_{\mathcal{X}} f(x)p(x)\,\mathrm{d}x,$$

and (partial) differentiation at a point x:

$$f \mapsto D_{x,i}[f] = \left.\frac{\partial f(z)}{\partial z_i}\right|_{z=x}.$$

Conditioning on linear functionals

It turns out that we can once again exploit the closure of the Gaussian distribution to linear transformations to condition a GP on f on the observation of any linear functional of f! This will allow us to both perform inference about f given observations of, for example, derivatives of f, and also to perform inference about linear functionals of f directly. This will provide us with a Bayesian mechanism for estimating integrals (a task traditionally called *quadrature*).

Suppose we have an unknown function $f\colon \mathcal{X} \to \mathbb{R}$ with the Gaussian process prior above:

$$p(f) = \mathcal{GP}(f; \mu, K),$$

and let L be a linear functional. We will write $\ell = L[f]$. Just as Gaussian distributions are closed under linear transformations, so are Gaussian processes closed under the evaluation of linear functionals! The prior distribution for ℓ is a Gaussian distribution:

$$p(\ell) = \mathcal{N}\bigl(\ell; L[\mu], L^2[K]\bigr)$$

where

$$L^2[K] = L\Bigl[L\bigl[K(\cdot, x')\bigr]\Bigr] = L\Bigl[L\bigl[K(x, \cdot)\bigr]\Bigr].$$

This result is essentially equivalent to the result for linear transformations of Gaussian-distributed vectors we have been using thus far, written with different notation. Notice also that if we consider the point-evaluation functional L_x, we recover a basic result:

$$p\bigl(f(x) \mid x\bigr) = \mathcal{N}\bigl(f(x); L_x[\mu], L_x^2[K]\bigr); = \mathcal{N}\bigl(f(x); \mu(x), K(x, x)\bigr).$$

Considering the integration functional, we obtain a perhaps more-interesting result:

$$p\biggl(\int f(x)p(x)\,\mathrm{d}x\biggr) = \mathcal{N}\biggl(\int f(x)p(x)\,\mathrm{d}x; \int \mu(x)p(x)\,\mathrm{d}x, \iint K(x, x')p(x)p(x')\,\mathrm{d}x\,\mathrm{d}x'\biggr).$$

Therefore a Gaussian process distribution on f implies a Gaussian distribution on its integral against an arbitrary function $p(x)$! Further, the problem of estimating the integral of the (perhaps quite complicated) function f has been reduced to the perhaps-simpler problem of integrating the mean and covariance functions μ and K. This is the main idea behind *Bayesian quadrature,* also called *Bayesian Monte Carlo.*

Given an observation of $L[f] = \ell$, we may condition our prior on this observation in a manner equivalent to that used to derive the posterior distribution of f. Let $\mathbf{X}$ be an arbitrary set of input locations. As before, we write the joint distribution between ℓ and $\mathbf{f} = f(\mathbf{X})$:

$$p\left(\begin{bmatrix}\mathbf{f}\\ \ell\end{bmatrix} \mid \mathbf{X}\right) = \mathcal{N}\left(\begin{bmatrix}\mathbf{f}\\ \ell\end{bmatrix}; \begin{bmatrix}\boldsymbol{\mu}\\ L[\mu]\end{bmatrix}, \begin{bmatrix}\mathbf{K} & ?\\ ? & L^2[K]\end{bmatrix}\right),$$

where we have defined:

$$\boldsymbol{\mu} = \mu(\mathbf{X}) \qquad \mathbf{K} = K(\mathbf{X}, \mathbf{X}).$$

To fill in the missing observations, we need to know the covariance between ℓ and the ith function value $f_i = f(\mathbf{x}_i)$. Here we can exploit the linearity of covariance:

$$\operatorname{cov}(f_i, \ell) = \operatorname{cov}\bigl(L_{\mathbf{x}_i}[f], L[f]\bigr) = L_{\mathbf{x}_i}\Bigl[L\bigl[\operatorname{cov}(f, f)\bigr]\Bigr] = L_{\mathbf{x}_i}\Bigl[L\bigl[K\bigr]\Bigr] = L\bigl[K(\mathbf{x}_i, \cdot)\bigr].$$

Now we have the general result

$$p\left(\begin{bmatrix}\mathbf{f}\\ \ell\end{bmatrix} \mid \mathbf{X}\right) = \mathcal{N}\left(\begin{bmatrix}\mathbf{f}\\ \ell\end{bmatrix}; \begin{bmatrix}\boldsymbol{\mu}\\ L[\mu]\end{bmatrix}, \begin{bmatrix}\mathbf{K} & L\bigl[K(\mathbf{X}, \cdot)\bigr]\\ L\bigl[K(\cdot, \mathbf{X})\bigr] & L^2[K]\end{bmatrix}\right).$$

Finally, we may condition this joint distribution on the observed value $\ell = L[f]$ to find the posterior of $\mathbf{f}$, which will be an updated multivariate Gaussian distribution. Because the set of points $\mathbf{X}$ was arbitrary, we may conclude that the posterior distribution is also a Gaussian process. The posterior mean and covariance functions are

$$\mu_{f|\ell}(\mathbf{x}) = \mu(\mathbf{x}) + \frac{L\big[K(\mathbf{x},\cdot)\big]}{L^2[K]}\big(\ell - L[\mu]\big);$$
$$K_{f|\ell}(\mathbf{x},\mathbf{x}') = K(\mathbf{x},\mathbf{x}') - \frac{L\big[K(\mathbf{x},\cdot)\big]L\big[K(\cdot,\mathbf{x}')\big]}{L^2[K]}.$$

We can easily extend this result to include multiple observations of functionals and also to incorporate Gaussian noise on each of these observations.

An example is shown in Figure 1, where we condition a Gaussian process prior on the integral observation $\int_0^{10} f(x)\,\mathrm{d}x = 5$. Notice that the posterior samples all have integral exactly equal to 5.

Bayesian Quadrature

Above, we conditioned a Gaussian process on an integral observation. In *Bayesian quadrature,* we do the opposite: given (potentially noisy) observations of a function $\mathcal{D} = (\mathbf{X}, \mathbf{y})$, we perform inference about an integral of interest, for example the expectation of f under a distribution p:

$$I_p[f] = \int f(x)p(x)\,\mathrm{d}x.$$

The traditional method for estimating integrals of this form is *Monte Carlo* estimation, where we sample some points $\{x_i\}_{i=1}^N$ from the distribution $p(x)$ and estimate

$$\int f(x)p(x)\,\mathrm{d}x \approx \sum_{i=1}^{N} f(x_i).$$

In Bayesian quadrature, we place a Gaussian process prior on f, which we condition on the observations $\mathcal{D}$. Notice that the input locations $\mathbf{X}$ do not need to be random samples from p, but rather we are allowed to evaluate f anywhere. The result is the posterior

$$p(f \mid \mathcal{D}) = \mathcal{GP}(f; \mu_{f|\mathcal{D}}, K_{f|\mathcal{D}}).$$

Following the above, we may also derive the posterior distribution of the expectation $I_p[f]$:

$$p\big(I_p[f] \mid \mathcal{D}\big) = \mathcal{N}\bigg(I_p[f]; \int \mu_{f|\mathcal{D}}(x)p(x)\,\mathrm{d}x, \iint K_{f|\mathcal{D}}(x,x')p(x)p(x')\,\mathrm{d}x\,\mathrm{d}x'\bigg).$$

For some choices of the prior prior mean and covariance functions μ and K and the distribution p, we may compute the required integrals exactly, giving a closed-form expression for the posterior distribution of the integral of interest.

Why is this useful? The main advantages to this approach are that we may explicitly model the structure of f via the covariance function K, and that the posterior variance of the integral may be used to derive an active sampling scheme, revealing the most-informative points to evaluate the function so as to estimate the integral with the highest precision. Note that the posterior variance of the integral only depends on where we sample the function, and not the actual values we observe. This property can be exploited to design optimal quadrature rules.

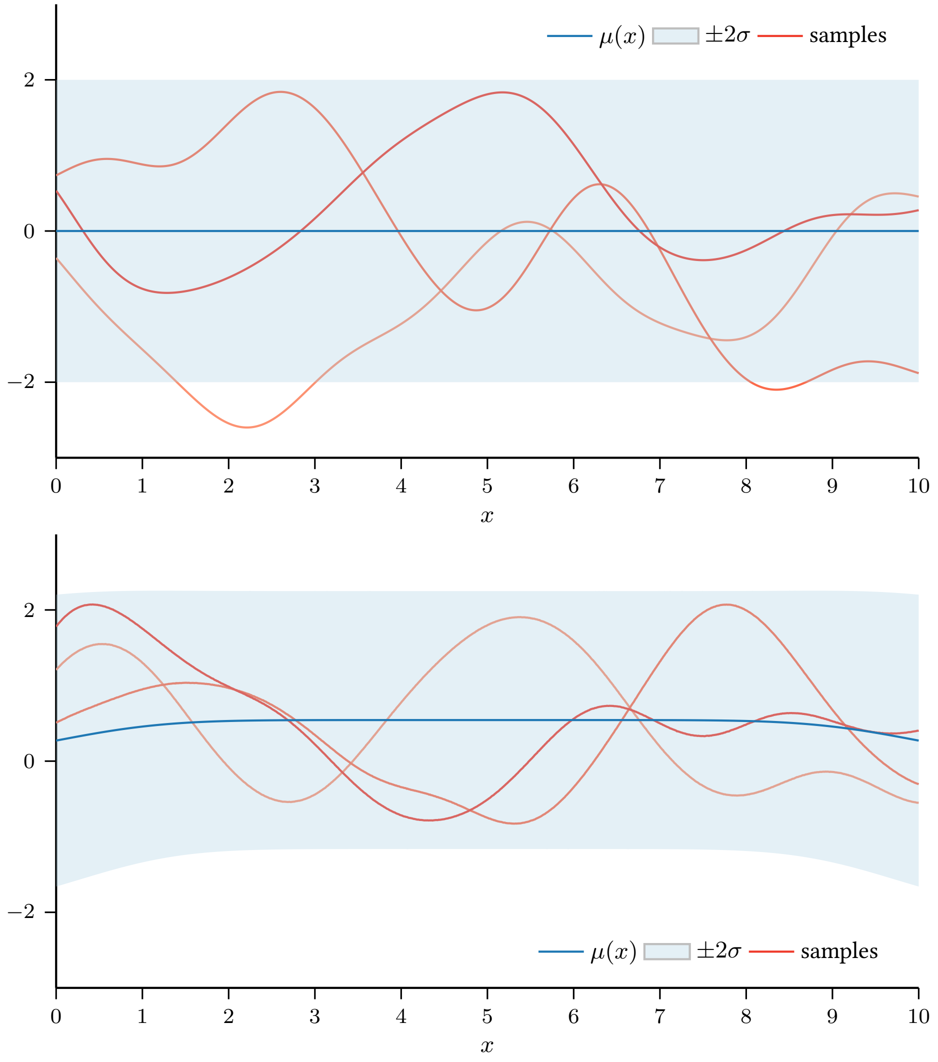

Figure 1: Above: a Gaussian process prior on a function f with mean zero and squared exponential covariance. Below: the posterior distribution on f after conditioning on the obesrvation $\int_0^{10} f(x)\,\mathrm{d}x = 5$. The posterior samples all have integral identically equal to 5.