Stochastic Process

a stochastic process is a collection of random variable indexed by some variable $x \in \mathcal{X}$. $f = \{ f(x) : x \in \mathcal{X} \}$ f can be thought of as a function of location x.

f is an infinite dimensional process. However, thankfully we only need consider the finite dimensional distributions (FDDs), i.e., for all $x_1, \dots x_n$ and for all $n \in N$ $P(f(x_1) \leq y_1, \dots, f(x_n) \leq y_n)$ as these uniquely determine the law of f .

A Gaussian process is a stochastic process with Gaussian FDDs, i.e. $(f(x_1) \dots f(x_n)) \sim \mathcal{N}(\mu, \Sigma)$

Import properties

- Property 1: $x \sim \mathcal{N}(\mu, \Sigma)$ if and only if $AX \sim \mathcal{N}_p(A\mu, A\Sigma A^T)$ So, sum of Gaussian is Gaussian (taking A as ones vector), and marginal distributions of multivariate are still Gaussian. (taking A is a vector with [0 0 0 1])
- Property 2: Conditional distributions are still Gaussian.

Suppose:

$$X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \mathcal{N}(\mu, \Sigma) \tag{20}$$

where,

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \tag{1}$$

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \tag{2}$$

Then when we only observe only the part of the result, here X_1:

$$(X_2 | X_1 = x_1) \sim \mathcal{N}(\mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(x_1 - \mu_1), \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}) \tag{3}$$

$$\begin{aligned} \pi\left(x_2 | x_1\right) &= \frac{\pi\left(x_1, x_2\right)}{\pi\left(x_1\right)} \propto \pi\left(x_1, x_2\right) \\ &\propto \exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right) \\ &\propto \exp\left(-\frac{1}{2}\left[\left(x_2-\mu_2\right)^\top Q_{22}\left(x_2-\mu_2\right) + 2\left(x_2-\mu_2\right)^\top Q_{21}\left(x_1-\mu_1\right)\right]\right) \\ &\propto \exp\left(-\frac{1}{2}\left[\left(x_2-\mu_2\right)^\top Q_{22}\left(x_2-\mu_2\right) + 2\left(x_2-\mu_2\right)^\top Q_{21}\left(x_1-\mu_1\right)\right]\right) \end{aligned}$$

$\text{So } X_2 | X_1 = x_1 \text{ is Gaussian.}$

$$\pi(x_2|x_1) \propto \exp\left(-\frac{1}{2}\left[(x_2-\mu_2)^\top Q_{22}(x_2-\mu_2)+2(x_2-\mu_2)^\top Q_{21}(x_1-\mu_1)\right]\right)$$
$$\propto \exp\left(-\frac{1}{2}\left[x_2^\top Q_{22}x_2 - 2x_2^\top(Q_{22}\mu_2+Q_{21}(x_1-\mu_1))\right]\right)$$
$$\propto \exp\left(-\frac{1}{2}\left(x_2-Q_{22}^{-1}(Q_{22}\mu_2+Q_{21}(x_1-\mu_1))\right)^\top Q_{22}\left(x_2-\ldots\right)\right)$$

So

$$X_2|X_1=x_1 \sim N(\mu_2+Q_{22}^{-1}Q_{21}(x_1-\mu_1), Q_{22})$$

A simple matrix inversion lemma gives

$$Q_{22}^{-1} = \Sigma_{22}-\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$$
$$\text{and} Q_{22}^{-1}Q_{21} = \Sigma_{21}\Sigma_{11}^{-1}$$

giving

$$X_2|X_1=x_1 \sim N\left(\mu_2+\Sigma_{21}\Sigma_{11}^{-1}(x_1-\mu_1), \Sigma_{22}-\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\right)$$

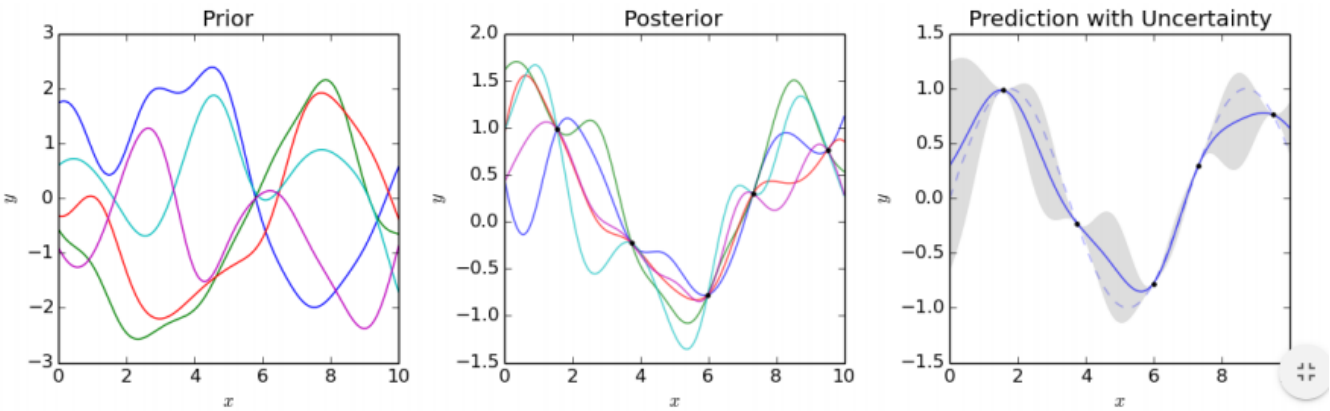
Conditional Update of GP

So suppose f is a Gaussian Process, then,

$$f(x_1), f(x_2)\ldots f(x_n), f(x) \sim \mathcal{N}(\mu, \Sigma) \tag{4}$$

if we observe its value at $x_1, \ldots, x_n$, then

$$f(x)|f(x_1)\ldots f(x_n) \sim \mathcal{N}(\mu^*, \sigma^*) \tag{5}$$



Why we use GP

1. The GP class of models is closed under various operations.
2. non-parametric/kernel regression
3. Naturalness of GP Framework
4. Uncertainty estimates from emulators

Why GPs: Closed under various operations

- Closed under addition $f1(\cdot), f2(\cdot) \sim GP$ then $(f1 + f2)(\cdot) \sim GP$
- Closed under Bayesian conditioning, i.e., if we observe $D = (f(x1), \ldots, f(xn))$ then $(f|D) \sim GP$
- Closed under any linear operation. If $f \sim GP(m(\cdot), k(\cdot, \cdot))$, then if L is a linear operator $L \circ f \sim GP(L \circ m, L^2 \circ k)$

Why GPs: Non-Parametric Regression

Suppose that we are given data

$$(x_i, y_i)_{i=1}^n \tag{6}$$

Linear Regression

$$y = x^T \beta + \epsilon \tag{7}$$

can be written in form of inner products

$$x^T x \tag{8}$$

$$\begin{aligned} \hat{\beta} &= argmin||y - X\beta||_2^2 + \sigma^2 ||\beta||^2 \\ \hat{\beta} &= (X^T X + \Sigma^2 I)^{-1} X^T y \\ \hat{\beta} &= X^T (XX^T + \sigma^2 I)^{-1} y \quad (the\ dual\ form) \end{aligned} \tag{9}$$

At first, the dual form looks like we've made the problem harder:

$$\begin{aligned} XX^T\ is\ n * n \\ X^T X\ is\ p * p \end{aligned} \tag{10}$$

but the dual form makes clear that linear regression only uses inner products.

The best prediction of y at a new location $x^{'}$ is:

$$\begin{aligned} \hat{y} &= x^{'T} \hat{\beta} \\ \hat{y} &= x^{'T} X^T (XX^T + \sigma^2 I)^{-1} y \\ \hat{y} &= k(x^{'})(K + \sigma^2 I)^{-1} y \end{aligned} \tag{11}$$

where,

$$k(x^{'}) := (x^{'} x_1, \ldots, x^{'} x_n) \tag{12}$$

and

$$K_{ij} := x_i^T x_j \tag{13}$$

these are kernel matrices, every element is the inner product between two rows of training points. And note the similarity to the GP conditional mean we derived before. If:

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \sim \mathcal{N}\left(0, \quad \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}\right) \tag{14}$$

then,

$$E(y^{'}|y) = \Sigma_{21}\Sigma_{11}^{-1}y \tag{15}$$

where, $\Sigma_{11} = K + \sigma^2 I$,and $\Sigma_{12} = \mathrm{Cov}(y, y')$ then we can see that linear regression and GP regression are equivalent for the kernel/covariance function $k\left(x, x'\right) = x^\top x'$

And we can replace the x by a feature vector in linear regression, e.g. $\phi(x) = (1 \ x \ x^2)$

Then

$$K_{ij} = \phi(X_i)^T \phi(X_j) \tag{16}$$

Generally, we don't think about these features, we just choose a kernel. But any kernel is implicitly choosing a set of features, and our model only includes functions that are linear combinations of this set of features (this space is called the Reproducing Kernel Hilbert Space (RKHS) of k).

Although our simulator may not lie in the RKHS defined by k, this space is much richer than any parametric regression model (and can be dense in some sets of continuous bounded functions), and is thus more likely to contain an element close to the true functional form than any class of models that contains only a finite number of features. **This is the motivation for non-parametric methods.** When we do GP, we are always assuming that the feature vector is in a much rich feature space.

Example

$$\phi(x) = \left(e^{-\frac{(x-c_1)^2}{2\lambda^2}}, \ldots, e^{-\frac{(x-c_N)^2}{2\lambda^2}}\right) \tag{17}$$

then as $N \to \infty$

$$\phi(x)^\top \phi(x) = \exp\left(-\frac{(x-x')^2}{2\lambda^2}\right) \tag{18}$$

Why GPs: Naturalness of GP framework

It has been shown, using coherency arguments, or geometric arguments, or..., that the best second-order inference we can do to update our beliefs about X given Y is

$$\mathbb{E}(X|Y) = \mathbb{E}(X) + \mathrm{Cov}(X, Y)\,\mathrm{Var}(Y)^{-1}(Y - \mathbb{E}(Y)) \tag{19}$$

i.e., exactly the Gaussian process update for the posterior mean. So GPs are in some sense **second-order** optimal. The **Second Order** means we only consider about the mean and the variance.

Why GPs: Uncertainty estimates from emulators

We often think of our prediction as consisting of two parts

- point estimate
- uncertainty in that estimate

Difficult of Using GPs

Kernel

We do not know what the covariance function, e.g. the kernel like. So what we can do is that we pick a covariance function from a small set, based usually on differentiability considerations. Possibly try a few (plus combinations of a few) covariance functions, and attempt to make a good choice using some sort of empirical evaluation.

Assume of the GPs

Assuming a GP model for your data imposes a complex structure on the data. The number of parameters in a GP is essentially infinite, and so they are not identified even asymptotically. So the posterior can concentrate not on a point, but on some submanifold of parameter space, and the projection of the prior on this space continues to impact the posterior even as more and more data are collected.

Hyper-parameter optimization

As well as problems of identifiability, the likelihood surface that is being maximized is often flat and multi-modal, and thus the optimizer can sometimes fail to converge, or gets stuck in local-maxima.